DOCUMENT RESUME

ED 147 344 TM 006 712

TITLE Proceedings of the Annual Conference of the Military Testing Association (18th, Gulf Shores, Alabama, October 18-22, 1976).

INSTITUTION Military Testing Association.; Naval Education and Training Program Development Center, Pensacola, Fla.

PUB DATE Oct 76

NOTE 1,457p.; Papers presented at the Annual Conference of the Military Testing Association (18th, Gulf Shores, Alabama, October 18-22, 1976) ; Best copy available

EDRS PRICE MF-$2.67 HC-$78.37 Plus Postage.

DESCRIPTORS Aptitude Tests; Computer Assisted Instruction; *Conference Reports; Criterion Referenced Tests; Educational Objectives; Enlisted Personnel; *Evaluation Methods; Females; *Job Analysis; Job Skills; *Military Personnel; Military Training; Occupational Tests; Officer Personnel; Performance Tests; *Personnel Evaluation; Personnel Selection; Predictive Ability (Testing); Promotion (Occupational); Test Bias; Test Construction; *Testing; Test Reliability; Test Validity; Vocational Aptitude

IDENTIFIERS Adaptive Testing; *Military Testing Association

ABSTRACT

The 75 papers included in these conference proceedings discuss testing conducted by the different brances of the armed forces. The importance of relating necessary job skills to the skills measured by the tests administered to the job applicants is emphasized. Various evaluation methods--including peer rating, aptitude testing, adaptive testing, performance or skill qualification testing, computer assisted testing, and job knowledge analysis--are used for personnel selection and evalaution regarding advancement. Additional topics discussed at the symposium include: the evaluation of military training programs, job satisfaction surveys, impact of female personnel in the military, and test construction. The By-laws of the Military Testing Association are appended. (GDC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED147344

18th Annual Conference of the Military Testing Association

PROCEEDINGS

BEST COPY AVAILABLE

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. L. BOYD

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Coordinated by

Naval Education and Training Program Development Center

Gulf Shores, Alabama

18-22 October 1976

2

TM006 712

18TH ANNUAL CONFERENCE
OF THE
MILITARY TESTING ASSOCIATION

MTA PRESIDENT

Captain Richard T. Thomas, USN

CONFERENCE COMMITTEE

CDR T. J. Cann
LT D. Harrison-Brown
M. A. Bachtel
S. D. Biggs
W. W. Birdsall
F. J. Conner

COORDINATED BY:

Naval Education and Training Program
Development Center, Ellyson
Pensacola, Florida 32509

Gulf State Park Motel Resort
Gulf Shores, Alabama
18 October - 22 October 1976

FOREWORD

The papers presented at the Eighteenth Annual Conference of the Military Testing Association came from the business, educational, and military communities, both foreign and domestic. The papers reflect the opinions of their authors only and are not to be construed as the official policy of any institution, government, or branch of armed service.

CONTENTS

Future Trends in Military Training– A Naval Perspective	VADM J. B. Wilson	1
Naval Enlisted Advancement System	M. J. Smith, Ph.D.	7
What is the Value of Aptitude Tests?	Raymond E. Christal	36
Staff Noncommissioned Officer Task Analysis Study	Major W. R. Masciangelo	52
Simulation of Combat as a Measure of Individual Proficiency	Robert Pulliam Boyd L. Mathers	80
Multiple Choice Tests: How Many Alternatives?	Major Ronald G. Swanson	104
Job Taxonomy/Duty Module Research for Officer Assignment Definition	Arthur L. Korotkin, Ph.D.	111
Attaining Equitable Personnel Allocations with Supply and Demand Quotas	Robert L. McMullen	126
Estimation of the Impartiality of Military Justice	Clark L. Hosmer, Ph.D.	135
Indirect Sampling of the Affective Domain to Assess "Petty Officer Qualities"	Raymond D. Bland, LCDR	142
Another Look at Peer Ratings	Major J. L. Mitchell Mr. Angelo DeNisi	150
The Chaplain Description Scale: A Peer Evaluation Without Tears	Chaplain (LTC) John S. Snyder Chaplain (LTC) D. R. Jernigan Chaplain (MAJ) Cecil D. Lewis	166
Reliability of Associate Ratings of Performance Potential by Army Aviators	Robert F. Eastman Robert L. McMullen	182
Total Testing in Coast Guard Aviation Training	Harry R. Anderson H. Kingsley Povenmire	193

Measurement of a Pilot's Control Skills in an Instrument Flight Maneuver	Brian D. Shipley, Jr. William V. Hagin Vernon S. Gerlach	202
Television as Stimulus Input in Synthetic Performance Testing Experiment I	John Cockrell	218
Sequential Analysis for Individual Proficiency Decisions	Claramae S. Knerr Kenneth I. Epstein	249
Introduction to the Symposium	Clay V. Brittain	270
The SQT Program as a Criterion-Referenced Approach to Proficiency Testing	Arthur G. Hermansen	272
Developing Skill Qualification Tests	William C. Osborn Roy C. Campbell	276
Establishing the Validity of SQT	Clay V. Brittain	292
Media Applications in the SQT Program	Cpt. Robert B. Wiltshire	298
The Institutionalization of Change in Military Training	Worth Scanland, Ph.D. Lawrence J. Korb Franklin F. Sands, Ph.D. Joseph Ward Gail T. Rayner R. M. Morgan Unna Huh Jeon	306
"The Other Ten Percent"–Revisited	J. E. Gerber, Jr., Ph.D.	353
An Air Force Job Satisfaction Research Project through 1976	R. Bruce Gould	368
A Computer Assisted Test Construction System	Vergil M. McIntosh	379
DYNA: Dynamic Programming Analysis and the Distribution of Effort	William J. Moonan, Ph.D.	389
Efficacies of Three Methods of Weighting Biodata to Predict Turnover	Warren S. Blumenfeld	425
Prediction of Test Outline Weights from Occupational Survey Data	David S. Vaughan, Capt	435

Computerized Adaptive Testing and its Potential for Military Testing Applications . . . . . James R. McBride 461

Estimation of Latent Trait Status Using Adaptive Testing Procedures . . . . . James B. Sympson 466

Adaptive Testing and the Problem of Classification . . . . . C. David Vale 488

Applications of Adaptive Testing in Measuring Achievement and Performance . . . . . Isaac I. Bejar 504

Applications of Item Response Theory to the Problem of Test Bias . . . . . Steven M. Pine 511

AFHRL Research on Utilization of Women in the Air Force . . . . . Jeffrey E. Kantor 521

The Utilization of Women in the Marine Corps . . . . . Steven Gorman 530

Research on the Aeronautical Adaptability of Women . . . . . Annette G. Baisden 541

A Segmentation Analysis of Women in the Army . . . . . Lawrence G. Corey 551

The Impact Number of Women in TOandE Units has on Mission Performance . . . . . Cecil D. Johnson 574

Content Validity for Training Analysis (A Practical Application) . . . . . LCDR C. Lee Walker 577

The Development of Statistically Parallel Tests by Analysis of Unique Variance . . . . . Malcolm James Ree 584

The Search for Fairness in Testing . . . . . Raymond L. Lewiski 590

Determination of Minimal Acceptable Performance Levels for Criterion-Referenced, Multiple-Choice Tests . . . . . John B. Meredith, Jr. Myron A. Robinson, Ph.D. 611

Discussion of a Study of Occupationally Transferable Skills . . . . . Allen A. Wiant 616

The Joint Services Approach to Enlistment Processing . . . . . William P. Acker 620

Problems of Billets Classification as seen by the Federal Armed Forces Association (DBwV) . . . . . Colonel H. E. Seuberlich 632

Occupational Analysis for Navy Instructional Systems Development (ISD): A Matrix Approach . . . . . D. D. Davis 662
T. M. Ansbro

Job Knowledge Analysis: A New Approach . . . . . Joseph S. Tartell, Capt. 706
Hendrick W. Ruck
Walter E. Driskill, Ph.D.

Job Knowledge Testing as a Measure of Training Effectiveness . . . . . Earl S. Stein 717

Computer-Assisted Occupational Counseling: A Prototype Model . . . . . Ted M. I. Yelien 743

The Selection of an Occupational Counseling Computer Network System for the Navy Recruiting Environment . . . . . Bernard A. Rafacz 765

Data Processing Problems in Collecting Job Survey Information . . . . . Johnny J. Weissmuller 783
Britt Kauffman

Relationships Between Attitudes Toward Computer-Assisted Instruction and Training Performance . . . . . Bruce W. Knerr 797

The Predictive Validity of Operational and Experimental Variables for Mental Group IVS in the Navy: A Review and Summary of the Findings From Four Sets of "Culture Fair" Tests . . . . . Charles H. Cory 812

Establishing Training Priorities in an Army Supply Military Occupational Specialty . . . . . Raymond O. Waldkoetter, Ed.D 836
Arthur C. F. Gilbert, Ph.D.
James L. Raney, Ph.D.
Helmut H. Hawkins

Development of CAI Performance Measures: TACFIRE Tactical Data System . . . . . W. G. Hoyt 847
A. K. Butler
P. W. Leung
J. E. Germas
J. T. Larson

Testing of Southeast Asia Refugees to Determine Military Enlistment	Major Wayne S. Sellman Captain Melvin B. Gambrell	865
Characteristics of Self Selected Job Content	Barry P. McFarland James B. Carpenter	890
Evaluation of G-Seat Cueing in Large Bodied Aircraft	John C. Griffin Barry P. McFarland Oak H. DeBerg	904
Prediction of Technical Training Success From ASVAB and Educational Background	Lonnie D. Valentine, Jr.	923
Prediction of Enlisted Tenure– (Two Years) The POET-2 Model	William A. Sands	955
Vocational Interest Assessment in the Air Force	William E. Alley	984
Evaluation of Five Food-Rating-Feedback Cards	Herbert Meiselman Day Waterman Nancy Cobean	1015
Worker Job Satisfaction and Opinion in Military Food Service System Design	Lawrence E. Symington	1045
Organizational Effectiveness Related to Non-Mos Task Performance of a Critical Subgroup	Raymond O. Waldkoetter, Ed.D. Arthur C. F. Gilbert, Ph.D. Helmut H. Hawkins Roy B. Wethy	1060
Evaluative Criteria for Comparing Models of Unit Performance	Donald E. Erwin	1070
Achievement Testing for Various Types of Concentrated Management Courses	John F. McAreavy, Ph.D.	1101
The Evaluation-in-Systems (EIS) Approach in Development of Training Programs	Maurice Dutton, Ph.D. Paul Liberty, Ph.D.	1117
Characteristics of an Officer Performance Evaluation Measure	Arthur C. F. Gilbert, Ph.D. Frances C. Grafton	1136

Performance Objectives: Problems in Definition Measurement, and Interpretation . . . . . Kenneth I. Epstein 1151

Score Quality Issues Related to Individual and Weapon Crew Criterion-Referenced Performance Tests . . . . . Frederick Steinheiser, Jr. 1166
C. Wesley Snyder, Jr.

Evaluation of Crew Gunnery Performance . . . . . G. Gary Boycan 1184
Andrew M. Rose

Evaluation of Junior Leadership Assessment Procedures . . . . . Captain Earle H. Bain 1197
Captain John B. Joaquin

From Task Inventories to Performance Appraisal Forms . . . . . Roy R. Van Cleve 1265

Ratio of Forces Techniques for Scoring Units in Combat Simulations . . . . . Dr. Edgar L. Shriver 1282

List of Conferees 1296

Military Testing Association, By-laws 1313

Index
Alphabetical listing—by Authors 1319

18TH MEETING

OF THE

MILITARY TESTING ASSOCIATION

19 OCTOBER 1976

VADM J. B. WILSON, USN

Good morning, ladies and gentlemen of the Military Testing Association. The Naval Education and Training Command and I as its Commander, are very pleased and proud to be your hosts for this, your 18th annual conference. Those of you who are "Plank Owners" as the Navy calls those who have been in the crew since the ship was commissioned--you plank owners will remember it was Navy Captain Harry GREAR (then Commanding Officer of the Naval Examining Center at Great Lakes, Illinois) who first proposed to the other services that an association of military evaluators would benefit all who might participate. Since that beginning the association has expanded its interests to embrace many other aspects of military training, and is the only professional forum designed exclusively for military trainers and evaluators. The opportunity to share with you some of my personal perspectives concerning military training is one I most happily accept, with the hope that in so sharing I might challenge you to look upward and outward toward improved ways to meet the requirements of the future, which are going to get increasingly more demanding and call for increasingly more initiatives from us all.

I have chosen to discuss with you the future trends in military training as we perceive them in the Navy for the reason that I view the current times to be the most likely since World War Two to challenge us to constructive and innovative change. We can take World War Two as a reference point because, at least for the Navy, those dark days marked the occasion of the major shift of individual skill training from shipboard to shore...from a responsibility of senior petty officers and junior officers of the ship's company to that of full time trainers in resident schoolhouses. In the ensuing thirty years this dependency upon professional and paraprofessional trainers to provide the combat units with fully trained personnel has not changed, and so it seems reasonable to select 1946 as a benchmark.

Since 1946 there have been significant changes in the American culture and society which have had major impacts upon the military, but these external changes have not brought forth correspondingly significant changes in the ways we train. Let me name a few of these changes.

First, for the first time in almost forty years we no longer have the draft to depend upon for a steady source of good quality personnel. Instead we must now take what we can get through the voluntary mechanism, and often what we can get is not exactly what we would choose. Yet we have not found ways in which to respond with precision and sensitivity to individual differences in trainability and we still push all hands through the training pipelines virtually without discrimination.

Certainly another change has been the outlook and motivation of today's youth. Thirty years ago young people coming of military age were accustomed, by and large, to a social environment requiring self-restraint and discipline, whereas the young people of today have been in too many cases brought up in an atmosphere of permissiveness, early release from parental controls, and open suspicion of, if not disdain for, those in authority. Yet statistics on the loss of first term personnel because of disciplinary problems clearly indicate that we have not achieved means by which to alter youthful attitudes through the training process.

Financial and promotion incentives have also changed. Thirty years ago a recruit in the military services began his enlistment with a monthly income of less than eighty dollars and therefore the incentive to aspire to higher rank and skill levels resided, among other places, in the sharp increase in income available through the process of promotion and career orientation. Today we pay a recruit more than three hundred and seventy-five dollars per month on top of all the many add-ons of housing, subsistence, medical care and so on, leaving very little room for incentives to increase income through high performance on the job. Yet we have not discovered alternative incentives to enhance the desire to achieve well in training and subsequently on the job.

As a final example of our failure to meet the challenges over the years, let me point out that the costs of training, as with those of all aspects of manpower management, have risen out of all proportion to the inflationary rate which has influenced all costs. Manpower, including the training of that resource, has now become more expensive than the military weapons systems it is manning, yet we have not developed ways by which we can achieve higher efficiency in training, that is, achieve more and better training for the investment required.

Now, what I have tried to do for the last few minutes is focus our attention on some of the more obvious changes which have taken place in the society, the culture and the economy since World War Two, while at the same time reminding us that we have been really quite ineffective in dealing with these changes. By the turn of the century will those of us still here have to confess to a continuing inability to respond to changing times and their impacts upon training, or will we say that we faced the changing times with imagination and creativity? That is the challenge I have in mind for us.

What are some of those changes which the future has in store for the military, and hence for military training? My crystal ball is no more accurate than yours, but some things seem to be fairly evident. A few of the changes I believe we can predict would include the following, for examples.

The trend has been with us for a long time, but I believe that in the coming years we will continue to be faced with ever more sophisticated weapons and sensors requiring ever more intensive training and skill by maintenance and operator personnel. A reliable evaluation by one of our service arms has disclosed a gap of over fifty percent between the capabilities of weapons systems in combat ready units and the abilities of the personnel to utilize those capabilities. That is clearly unacceptable from an economics standpoint, but far more important, it is totally unacceptable from a readiness standpoint. The numerical weapons superiority of our possible enemy is of such a degree that we must be able to get every ounce of potential effectiveness from every weapon in the field. Only training can produce that condition, and obviously the training we are currently delivering is not. Better training both in resident schoolhouse sites and follow-on in the field is a tremendous challenge for us all.

Youngsters who are accustomed to a philosophy that teaches that someone, government or someone else, will provide for the necessities of life are not inclined to see the advantages of self-development through taking advantage of training opportunities available to them. Too many, especially those who have been through traumatic school experiences involving failure or disenchantment with the school system, are not interested in taking full advantage of the schools we offer them in the military. How to identify these people early in their enlistments, and how to motivate them to a sense of self-achievement is a challenge worthy of the best of our learning psychologists. I consider this a major problem for the future, inhibiting the development of the kind of career oriented military posture we require.

And then there is the problem of identifying and attending to individual differences between people, a problem very close to the hearts of test and measurement experts. If we are indeed losing over the one third of our first termers because of disciplinary or other problems of maladjustment, we obviously are not able to discriminate at accession points between the potential achiever and the potential dropout. It is a challenge of transcendental significance that we develop tests and measures which can discriminate among people in the kinds of ways which we require. It is economically catastrophic to invest training in those who cannot succeed, and evaluators must get with it and provide us with discriminators of high validity and reliability.

The military services of the future will become, certainly, less and less of a career force unless we can find means by which to reduce the loss of our first term enlistees, and this may prove a problem without a solution. The Army, the Navy, the Air Forces, the Marine Corps will continue, until such time as our national policy is dramatically changed, to be deployed around the world, often in unwelcoming and alien environments. In comparison with careers where families can find stability and a sense of continuity, the military

cannot compete. Therefore, it is my opinion that we will have to develop manpower policies and practices, including training policies and practices, which take into account that we will always have to deal with a mind-boggling annual turnover of personnel. In fiscal year 1975, for example, the aggregate turnover rate in the fleet was 56 percent. Such a level of turnover drives training cost sky-high, and therefore the challenge to us trainers is to devise means for training which somehow adjust to the challenge of equipment sophistication on the one hand yet attend to the requirement for minimum training investments in non-career personnel on the other. That is no mean challenge indeed, yet we are duty bound to meet it or see the end of professionalism in our military forces.

Finally, although I will have left many challenges unmentioned in this short discussion, let me mention the future of resident schoolhouse training as I perceive it. My personal interpretations of the decisions coming from the congressional armed services committees is that they cannot and will not forever accept the rising costs of personnel in the services, to which training and education is a major contributor. We have already seen the inroads being made in educational programs in the military...programs which we believe are so important to develop military personnel with the wisdom to make good decisions. I expect to see in the near term years similar inroads into our resident school individual skill training programs, despite our loud protestations that such skill training is the key to weapons systems readiness. The mandate, I am afraid, is going to be to find other, less expensive ways by which to have personnel acquire these necessary skills, and I leave you with yet another challenge for the future to come up with means by which to accomplish that end.

I am not by any yardstick a pessimist. To the contrary, I view the future with optimism, if somewhat strained, and with excitement, unconstrained. If challenge is the adrenalin required for achievement worthy of the name, and I believe it is. Then we are in for some exciting times ahead. But let me be sure you share with me the conviction that it will take all the imagination, all the ingenuity, all the hard work, all the willingness to stick our necks out by suggesting different ways to do the job, all of our combined professionalism, to meet the challenges of the future if they are to be met. Of course the alternative is to just keep on doing things in the same old ways in which we feel comfortable and at ease. But I trust you will not take that course, and I am therefore confident that it can be said of the members of this association that they will have the perception to see the need for changes where such a need does indeed exist, they will have the skills and inventiveness to discover the right ways to change, and they will have the courage to make needed changes in the face of the absolutely predictable opposition from those who won't want to rock the boat.

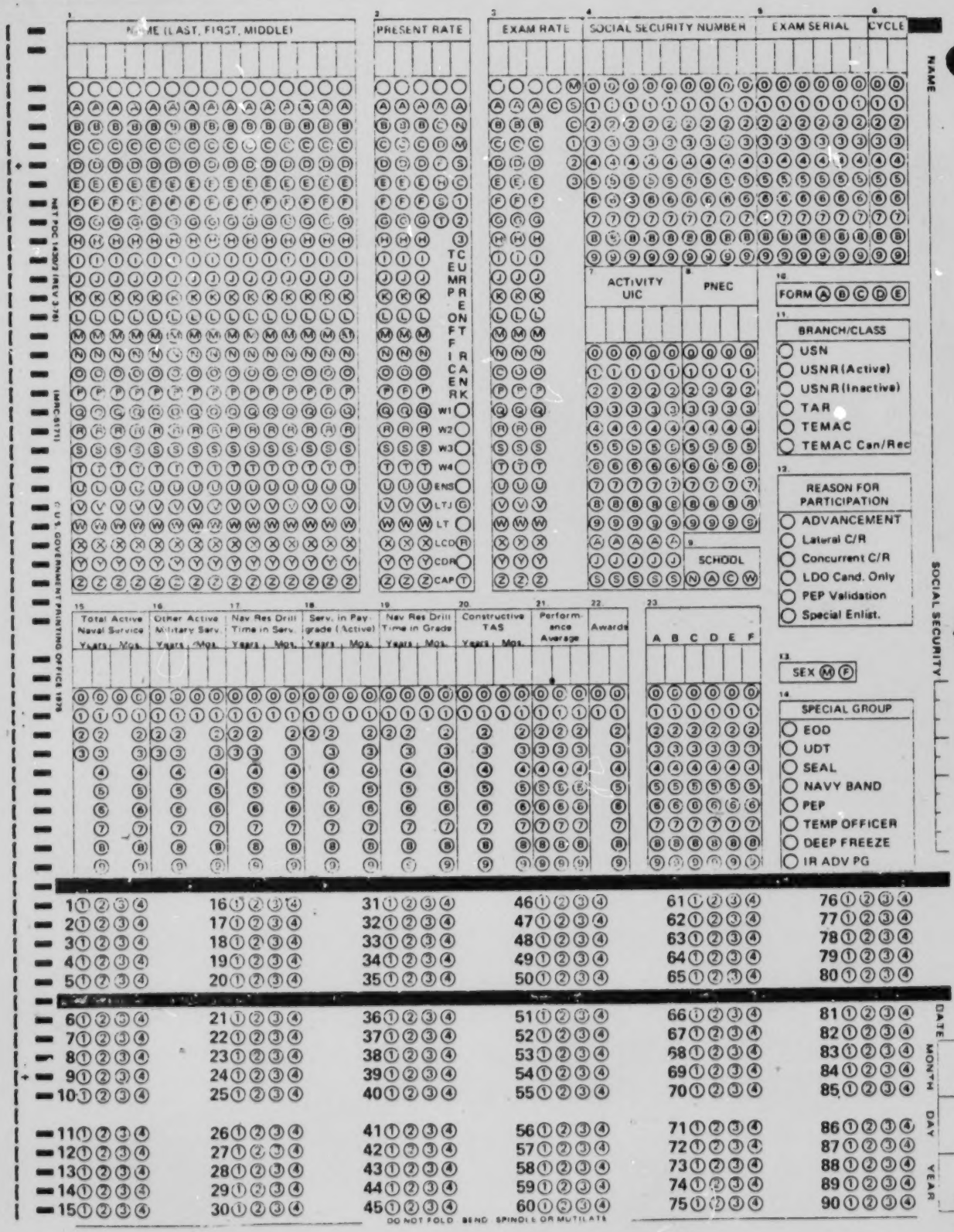
I appreciate very much the opportunity to speak with you this morning and to help in the launching of your eighteenth annual conference. I wish you the best of good debate and discussion, and look forward to the report of your deliberations.

Naval Enlisted Advancement System
M. J. SMITH, Ph.D.
NETPDC, Pensacola, Florida

I would like to describe the Advancement System currently in use for Navy enlisted personnel. For E4 through E6, examinations are administered twice each year, and for E7 through E9, once a year. Before a candidate can take an examination certain prerequisites are required. He must have at least a given amount of time in service and time in rate dependent upon the pay grade. He must have completed a check-off list of performance tasks which are different for each rating and pay grade. And finally, the candidate is required to have the recommendation of his Commanding Officer that he is fully qualified to perform in the rating at the next higher pay grade. The advancement tests we give are developed to measure how well the candidate knows the materials and tasks of the pay grade to which he is trying to be advanced.

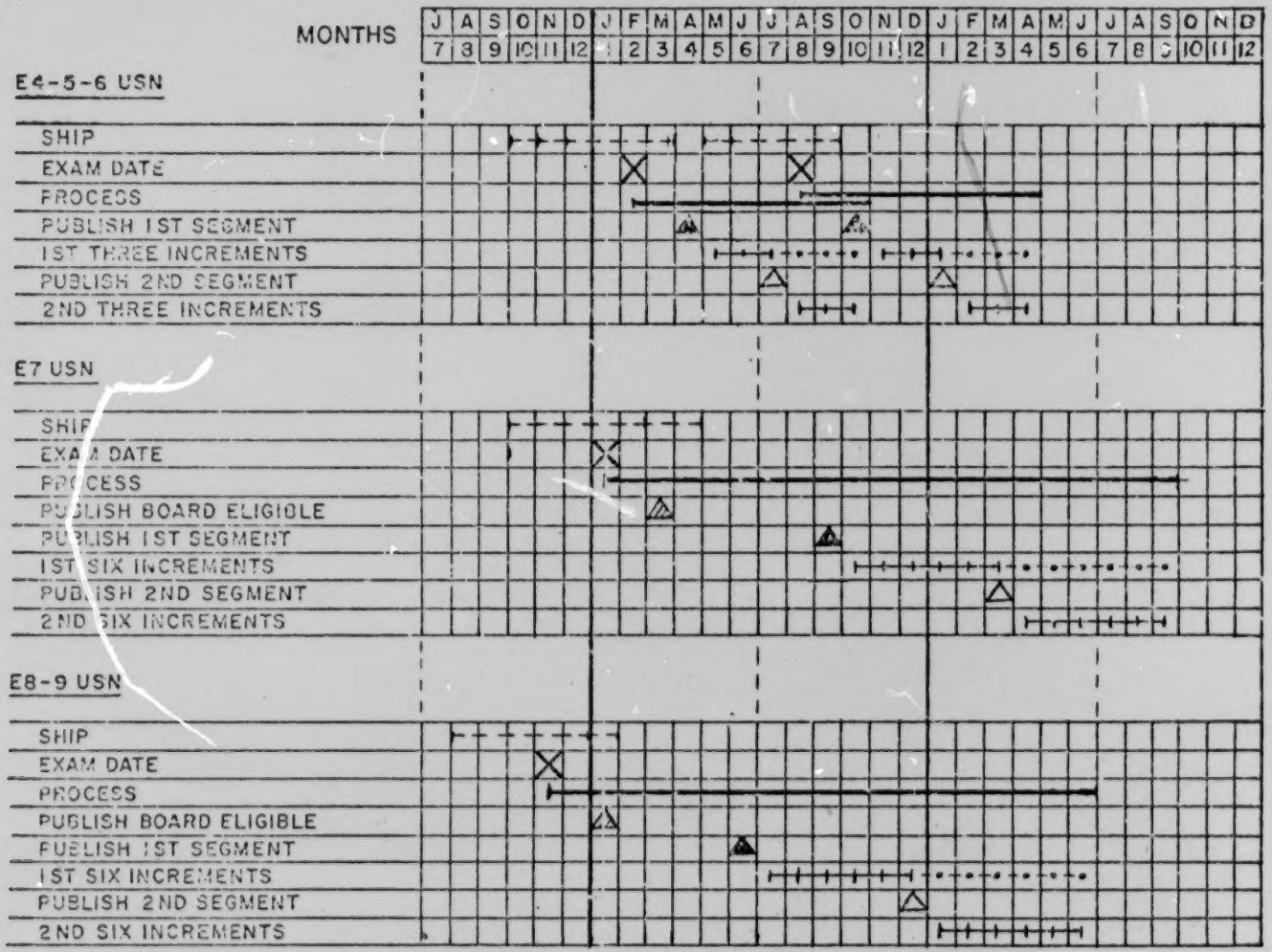
Prior to the date of the exam, a work sheet is prepared which has the person's name, social security number, their present rate and their exam rate, and all other information concerning the candidate which is required. On the day of the exam the candidates copy all of this information from their worksheets onto the answer sheets (shown in Slide 1). They then answer 150 multiple choice test items within a three hour time limit.

Slide 2 shows the various phases of the exam cycles of the Navy Enlisted Advancement System. During the month of February, usually on a Tuesday, all E4's in the Navy take the exams, two days later E5's take the exams, and on the Tuesday of the next week E6's take the exams. Actually, there are approximately 90 unique exams for each pay grade for a total of


1 NAME (LAST, FIRST, MIDDLE)
2 PRESENT RATE
3 EXAM RATE
4 SOCIAL SECURITY NUMBER
5 EXAM SERIAL
6 CYCLE
7 ACTIVITY UIC
8 PNEC
9 SCHOOL
10 FORM A B C D E
11 BRANCH/CLASS
USN
USNR (Active)
USNR (Inactive)
TAR
TEMAC
TEMAC Can/Rec
12 REASON FOR PARTICIPATION
ADVANCEMENT
Lateral C/R
Concurrent C/R
LDO Cand. Only
PEP Validation
Special Enlist.
13 SEX M F
14 SPECIAL GROUP
EOD
UDT
SEAL
NAVY BAND
PEP
TEMP OFFICER
DEEP FREEZE
IR ADV PG
15 Total Active Naval Service Years Mos.
16 Other Active Military Serv. Years Mos.
17 Nav Res Drill Time in Serv. Years Mos.
18 Serv. in Pay-grade (Active) Years Mos.
19 Nav Res Drill Time in Grade Years Mos.
20 Constructive TAS Years Mos.
21 Performance Average
22 Awards
23 A B C D E F
NAME
SOCIAL SECURITY
DATE MONTH DAY YEAR
DO NOT FOLD BEND SPINDLE OR MUTILATE
NET POC (43007 (REV 3-76)
(NRC-61171)
U.S. GOVERNMENT PRINTING OFFICE 1978

SLIDE 1

NAVY ENLISTED ADVANCEMENT SYSTEM


MONTHS
J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D
7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12
E4-5-6 USN
SHIP
EXAM DATE
PROCESS
PUBLISH 1ST SEGMENT
1ST THREE INCREMENTS
PUBLISH 2ND SEGMENT
2ND THREE INCREMENTS
E7 USN
SHIP
EXAM DATE
PROCESS
PUBLISH BOARD ELIGIBLE
PUBLISH 1ST SEGMENT
1ST SIX INCREMENTS
PUBLISH 2ND SEGMENT
2ND SIX INCREMENTS
E8-9 USN
SHIP
EXAM DATE
PROCESS
PUBLISH BOARD ELIGIBLE
PUBLISH 1ST SEGMENT
1ST SIX INCREMENTS
PUBLISH 2ND SEGMENT
2ND SIX INCREMENTS

SLIDE 2

approximately 270 unique exams given in February. In August 270 new unique exams are administered, which means about 540 exams are written and administered for pay grades E4 through E6. In January on a given day all E7 candidates take exams, which number approximately 85 unique exams. In November on a given day all E8 and E9 candidates take exams, which amount to some 170 unique exams.

Approximately four months prior to each exam month we ship an ordering letter to each Navy activity. This letter contains a list of the exams to be given for the next cycle and an ordering form. The ordering form is an Optical-Scan type document. The activities mail the form to us indicating the number of each type exam which they plan to administer on the upcoming exam dates. After receiving this form back, we scan the ordering form using a Sentry 70, which is then processed through an IBM 360 resulting in a shipping list of the exams ordered, each identified by a serial number. The required number of exams and answer sheets are then shipped to the activity requesting them. This phase is shown as a dashed line on Slide 2.

Within a few days after the day of the examination, we start receiving the answer sheets back in Pensacola. For approximately one and a half months after the exam we continue to receive answer sheets, which are scanned and all the information is stored on computer tape. This is the beginning of the processing phase shown as a solid line on Slide 2.

When we estimate that about 95 percent of the answer sheets of exam-takers have been returned, we score the exams using keys for these exams stored in the computer. (At present we are using a Univac 418 computer but are in the process of converting to an IBM 360/65.) The raw scores for each exam, which we speak of as a competing group, are standardized

with a mean standard score of 50 and a standard deviation of 10, with a minimum of 20 and a maximum of 80. In addition to how well a candidate does on an advancement examination the individual's performance evaluation average in his present rate counts toward advancement. For E4 through E6 candidates, credit is also given for some of their experience which will be discussed later. For each candidate a Final Multiple Score, referred to as his FMS, is calculated based on his exam standard score, performance average, and for E4 through E6 candidates his experience credits. Within each exam competing group the candidates are rank ordered on the basis of highest FMS to the lowest FMS.

For pay grades E4 through E6, we send to BUPERS the number of candidates who have passed the exam in each rate, and they send back to us the number of individuals required by rating and pay grade to meet the manning needs of the Navy for a six month period. The number required in each rate is counted down the FMS ranking, setting a minimum FMS-cut. Each candidate whose FMS is equal to or greater than the minimum FMS-cut for his given rate is considered to be a selectee for advancement during the six months of advancements for that exam cycle. We also receive from BUPERS the number of advancements per month by pay grade they require for the first of two segments, each of three monthly increments. The three months of advancements are fair-shared among the selectees within a pay grade. When we publish to the fleet the first segment (shown as solid triangles on Slide 2), candidates are authorized to be advanced on the sixteenth of each of the three months and the status of the remaining candidates are indicated as selectee, pass-not-advanced, or fail. When we publish the second segment (shown as open triangles on Slide 2), the remaining selectees

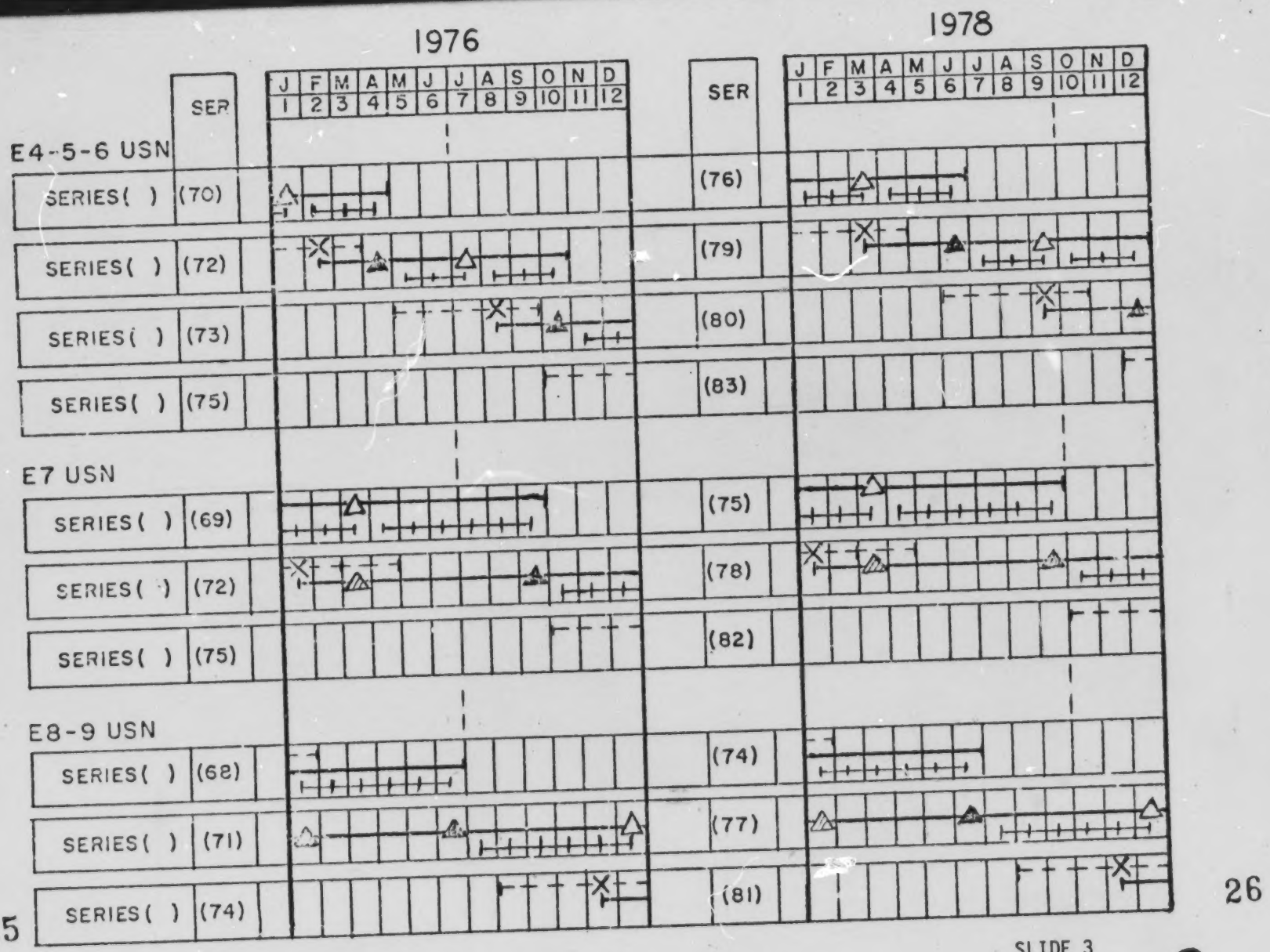
are authorized to be advanced on the sixteenth of one of the three months in the second segment. FMS-cuts are established for each monthly increment within each competing exam group; those candidates with the highest FMS's are advanced first and those with FMS's equal to or just above the minimum FMS cut are advanced last within an exam cycle.

Candidates for pay grades E7 through E9 are selected for advancement by selection board action. The candidates whose records are reviewed by one of these boards are chosen on the basis of their FMS which includes only their exam standard score and performance average. Again the FMS's within a competing exam group are rank ordered from high to low. We send the names of the top 50 percent of the E7's and the top 75 percent of the E8's and E9's to a selection board and publish to the fleet the status of all candidates as selection-board-eligible, selection-board-ineligible, and fail. This publication is shown as a triangle enclosing parallel lines on Slide 2. After a selection board adjourns, we receive the names of the candidates selected for advancement with selectee numbers assigned by the board to persons within each rate. We, also again, receive from BUPERS the number of advancements per month by paygrade they require for the first of two segments, each of which are six monthly increments for E7 through E9. The six months of advancements are fair-shared among the selectees within a paygrade. When we publish to the fleet the 1st segment (shown as solid triangles on Slide 2), board eligible candidates are either authorized advancement in one of the first six increments or indicated as selectees or non-selectees. When we publish the 2nd segment (shown as open triangles on Slide 2), the remaining selectees are authorized advancements in one of the six months in the second segment. Rather than FMS-cuts, Selectee-number-

group; thus candidates selected by a board with the lowest selectee-numbers are advanced first and those with the higher selectee-numbers are advanced last within an exam cycle.

As may be noted in Slide 2, the exam cycles extend over more than one calendar year. This results in the need for us to track and publish several cycles for each of the pay grades during a given calendar year. After each 1st segment notification, weekly update publications are made involving discrepancy corrections, activity changes and late returns. Slide 3 shows the cycles/series we processed during 1976 (using the same symbols as Slide 2): four for E4-6, three for E7 and three for E8-9. Also shown are the planned cycles/series for 1978. As may be noted, the E4-6 exam dates have been moved one month later and the advancement increments two months later. This change was occasioned by the new fiscal year accounting beginning in October rather than July. As may be seen, the publication dates (shown as triangles) will tend to be grouped in the same months in 1978 as compared to 1976.

In Slide 4, the computation equations for FMS's at different pay grades are given. In 1974 the Chief of Naval Personnel decided that we were advancing people who were good at taking exams but perhaps missing people who were good performers but poor at taking exams. A policy decision was published stating what each FMS component should contribute percentage-wise to the FMS. For example at E6, the exam was to count 30 percent, performance 35 percent, and experience 35 percent. Using a maximum-point weighting model the equations shown were developed. The key to these equations is the fact


1976
1978
J F M A M J J A S O N D
1 2 3 4 5 6 7 8 9 10 11 12
SER
E4-5-6 USN
SERIES( ) (70)
SERIES( ) (72)
SERIES( ) (73)
SERIES( ) (75)
E7 USN
SERIES( ) (69)
SERIES( ) (72)
SERIES( ) (75)
E8-9 USN
SERIES( ) (68)
SERIES( ) (71)
SERIES( ) (74)
(76)
(79)
(80)
(83)
(75)
(78)
(82)
(74)
(77)
(81)

25

26

SLIDE 3

NAVAL ENLISTED ADVANCEMENT COMPUTATION EQUATIONS

Pass/Fail examination raw score (RS) cuts:

E4 RS min = 43 E5 RS min = 48 E6-9 RS min = 53

. Final multiple score (FMS) of candidates who PASS

Paygrade	Factor		Computation		Max Points	%
E4/E5	Examination:	Std Scr(SS)			80	35
	Performance:	(Min 3pt)	60(Max 4)-170		=70	30
	Experience:					35
	LOS-TIR		(Max 15 yrs)+15		=30	(13)
	TIR		2(Max 7½ yrs)+15		=30	(13)
	AWD		(Max 10 pts)		=10	(4.5)
	PNA		(Max 10 pts)		=10	(4.5)
				Max FMS	230	100
E6	Examination:	Std Scr(SS)			80	30
	Performance:	(Min 3pt)	10(Max 9)+2		=92	35
	Experience:					35
	LOS-TIR		(Max 15 yrs)+19		=34	(13)
	TIR		2(Max 7½ yrs)+19		=34	(13)
	AWD		(Max 12 pts)		=12	(4.5)
	PNA		(Max 12 pts)		=12	(4.5)
				Max FMS	264	100
E7	Examination:	Std Scr(SS)			80	60
	Performance:		6(Max 9)-2		=52	40
				Max FMS	132	100
E8	Examination:	Std Scr(SS)			80	50
	Performance:		9(Max 9)-1		=80	50
				Max FMS	160	100
E9	Examination:	Std Scr(SS)			80	40
	Performance		15(Max 9)-15		=120	60
				Max FMS	200	100

I. Lateral change of rate standard score (SS) cut for E4-9 is 37 min.

27 SLIDE 4

that we standardize all exams to the typical Navy scale of a minimum of 20, a mean of 50, and a maximum of 80 points. It can, perhaps, be most easily illustrated at E8 where the desired percentages are an even 50-50 split between the examination and performance. Thus, given that the maximum number of points for the exam is 80, the maximum performance must also be equal to 80 points. Performance evaluations at E8 are based on a 1 to 9 point subjective rating scale, thus multiplying 9 times the 9-point maximum gives 81 and subtracting 1 gives the desired equation for performance with a maximum of 80 points and a maximum FMS of 160 at E8. At the other pay grade, it may be noted, that the exam maximum of 80 is also the desired percentage of the maximum FMS for that pay grade, and all of the equations shown for the other components result in maximum points equal to the desired percentage of the maximum FMS at the same pay grade.

The next group of slides show typical raw score distributions of the FMS components at pay grade E6. The data is from a sea-points study we did of Series 69/Feb '75 E6 male candidates, who number 16,329.

Slide 5 shows the examination raw scores. In this sample the low score was 28 and the high score was 131 with a mean of 73.5. It usually is a surprise to people that the candidates do not score higher on 150-item tests. The exams are written by our Navy Chiefs in their own specialties, and it should be remembered that the tests are designed to measure, primarily, knowledge required at the next higher pay grade than the candidates are presently in.

Slide 6 shows the distribution of performance averages transformed to a one to four point scale. The low average in this sample was 3.01 and the

NETPDC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6) VB. 40 EXAM RAW SCORE 40

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(103.000)
16329	73.5205	15.2617	232.9184	0.2782	-0.1626	28.000 TO 131.000

LOWEST EXACT LIMIT = 27.50000 INTERVAL = 3.00000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
28.000	30.000	3	3	0.018	0.018	X
31.000	33.000	9	12	0.055	0.073	X
34.000	36.000	16	28	0.098	0.171	X
37.000	39.000	36	64	0.220	0.392	X
40.000	42.000	92	156	0.563	0.955	XXXX
43.000	45.000	191	347	1.170	2.125	XXXXXX
46.000	48.000	276	623	1.690	3.815	XXXXXXXXXX
49.000	51.000	465	1088	2.848	6.663	XXXXXXXXXXXXXXXX
52.000	54.000	594	1682	3.638	10.301	XXXXXXXXXXXXXXXXXXXXX
55.000	57.000	772	2454	4.728	15.029	XXXXXXXXXXXXXXXXXXXXXXXXX
58.000	60.000	882	3336	5.401	20.430	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
61.000	63.000	1078	4414	6.602	27.032	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
64.000	66.000	1205	5619	7.380	34.411	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
67.000	69.000	1204	6823	7.373	41.785	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
70.000	72.000	1289	8112	7.894	49.678	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
73.000	75.000	1237	9349	7.575	57.254	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
76.000	78.000	1114	10463	6.822	64.076	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
79.000	81.000	1036	11499	6.345	70.421	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
82.000	84.000	942	12441	5.769	76.190	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
85.000	87.000	890	13331	5.450	81.640	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
88.000	90.000	720	14051	4.409	86.049	XXXXXXXXXXXXXXXXXXXXXXXXXX
91.000	93.000	617	14668	3.779	89.828	XXXXXXXXXXXXXXXXXXXX
94.000	96.000	480	15148	2.940	92.767	XXXXXXXXXXXXXXXXX
97.000	99.000	314	15462	1.923	94.690	XXXXXXXXXXX
100.000	102.000	255	15717	1.562	96.252	XXXXXXXXX
103.000	105.000	197	15914	1.206	97.459	XXXXXXX
106.000	108.000	147	16061	0.900	98.359	XXXXX
109.000	111.000	92	16153	0.563	98.922	XXX
112.000	114.000	68	16221	0.416	99.339	XXX
115.000	117.000	59	16280	0.361	99.700	XX
118.000	120.000	24	16304	0.147	99.847	X
121.000	123.000	19	16323	0.116	99.963	X
124.000	126.000	4	16327	0.024	99.988	X
127.000	129.000	1	16328	0.006	99.994	X
130.000	132.000	1	16329	0.006	100.000	X

17

SLIDE 5

NETPDC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6) VB. 41 FOUR POINT AVG 41

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(0.990)
16329	3.6622	0.1280	0.0164	-0.1948	0.2004	3.010 TO 4.000

LOWEST EXACT LIMIT = 2.97500 INTERVAL = 0.03000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
3.010	3.030	1	1	0.006	0.006	X
3.040	3.060	5	6	0.031	0.037	X
3.070	3.090	7	13	0.043	0.080	X
3.100	3.120	2	15	0.012	0.092	X
3.130	3.150	1	16	0.006	0.098	X
3.160	3.180	2	18	0.012	0.110	X
3.190	3.210	5	23	0.031	0.141	X
3.220	3.240	1	24	0.006	0.147	X
3.250	3.270	16	40	0.098	0.245	X
3.280	3.300	39	79	0.239	0.484	X
3.310	3.330	49	128	0.300	0.784	XX
3.340	3.360	90	218	0.551	1.335	XXX
3.370	3.390	147	365	0.900	2.235	XXXXX
3.400	3.420	224	589	1.372	3.607	XXXXXX
3.430	3.450	343	932	2.101	5.708	XXXXXXXXXX
3.460	3.480	493	1425	3.019	8.727	XXXXXXXXXXXXXXX
3.490	3.510	706	2131	4.324	13.050	XXXXXXXXXXXXXXXXXXXXX
3.520	3.540	897	3028	5.493	18.544	XXXXXXXXXXXXXXXXXXXXXXXXX
3.550	3.570	1118	4146	6.847	25.390	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3.580	3.600	1330	5476	8.145	33.535	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3.610	3.630	1445	6921	8.849	42.385	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3.640	3.660	1539	8460	9.425	51.810	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3.670	3.690	1467	9927	8.984	60.794	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3.700	3.720	1426	11353	8.733	69.527	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3.730	3.750	1227	12580	7.514	77.041	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3.760	3.780	1077	13657	6.596	83.636	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
3.790	3.810	841	14498	5.150	88.787	XXXXXXXXXXXXXXXXXXXXXXXX
3.820	3.840	653	15151	3.999	92.786	XXXXXXXXXXXXXXXXXX
3.850	3.870	484	15635	2.964	95.750	XXXXXXXXXXXXXX
3.880	3.900	325	15960	1.990	97.740	XXXXXXXXX
3.910	3.930	203	16163	1.243	98.983	XXXXXX
3.940	3.960	101	16264	0.619	99.602	XXX
3.970	3.990	54	16318	0.331	99.933	XX
4.000	4.020	11	16329	0.067	100.000	X

18

32

SLIDE 6

31

high average was 4.00 with a mean of 3.66. What is perhaps most interesting about this distribution is its normal shape--if one starts at about 3.25 and goes up to 4.00. E6 candidates are required to have a minimum of 3.00 in order to be recommended.

Slide 7 shows the distribution of length of service in years. The low point was 2.58 years and the high point was 30.17 years with a mean of 8.78 years. In this highly skewed distribution the peak occurred in the six to seven interval, and almost 50% of the candidates had only five to eight years of service.

Slide 8 shows the distribution of time in rate in years. The low point was 0.58 years and the high point was 20.67 years with a mean of 4.34 years. Again the distribution is highly skewed with 30% of the candidates having between 2.25 and 3.25 years in their present rate.

Slide 9 shows the distribution of award points. Different medals and awards count different points. In this sample 8½% of the candidates had no awards and one person had a credit for 19 points, and the mean was 3.03 points. A little more than 50% of the candidates had two points to their credit.

Slide 10 shows the distribution of PNA points, which stands for Passed-Not-Advanced. In each exam cycle if a candidate passes the exam but is not advanced he may earn PNA points based upon his standing on the exam and on his performance evaluations. If the candidate was in the top quarter of either distribution he receives 1.5 points, in the second upper quarter 1.0 point, and in the third upper quarter 0.5 point. Thus it is possible to

NETPDC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6) VB. 42 LOS YRS 42

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(	27.583)
16329	8.7791	3.4343	11.7947	1.2073	1.1143	2.583 TO	30.167

LOWEST EXACT LIMIT = 1.99500 INTERVAL = 1.00000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
2.000	2.990	2	2	0.012	0.012	X
3.000	3.990	23	25	0.141	0.153	X
4.000	4.990	662	687	4.054	4.207	XXXXXXXXX
5.000	5.990	2618	3305	16.033	20.240	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
6.000	6.990	3156	6461	19.328	39.568	XXXXXXXXXXXXXXXXXXXX

NETPDC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6) VB. 44 TIR YRS 44

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(20.083)
16329	4.3366	1.9840	3.9363	1.2424	2.1592	0.583 TO 20.667

LOWEST EXACT LIMIT = -0.25500 INTERVAL = 0.50000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
0.250	0.740	1	1	0.006	0.006	X
0.750	1.240	3	4	0.018	0.024	X
1.250	1.740	15	19	0.092	0.116	X
1.750	2.240	1242	1261	7.606	7.722	XXXXXXXXXXXXXXXXXXXXX
2.250	2.740	2467	3728	15.108	22.831	XXXXXXXXXXXXXXXXXXXX

NETPDC SEA PT STUDY--SERIES 69-FEB 75-E6 MALES N=16329 (98.54% OF E6) VB. 45 AWD NUM 45

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(19.000)
16329	3.0333	2.1110	4.4562	1.7115	5.1573	0.000 TO 19.000

LOWEST EXACT LIMIT = -0.50000 INTERVAL = 1.00000 NUMBER OF CASES? WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
0.000	0.000	1385	1385	8.482	8.482	XXXXXXX
1.000	1.000	80	1465	0.490	8.972	X
2.000	2.000	8372	9837	51.271	60.243	XXXXXX

NETPDC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6) VB. 46 PNA PTS 46

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(14.000)
16329	1.6390	2.2347	4.9940	1.6763	2.7022	0.000 TO 14.000

LOWEST EXACT LIMIT = -0.05000 INTERVAL = 0.50000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
0.000	0.400	7782	7782	47.658	47.658	XXXX

obtain 3.0 PNA points in each exam cycle, and it is permitted to accumulate these points over the past five exam cycles. In this sample almost 50% had received no PNA points from past cycles but one person had accumulated 14 points. The mean was 1.64 points.

The next three slides show the distribution of exam, performance, and experience points after the application of the E6 computation equations of Slide 4 to the raw data of the preceding distributions.

Slide 11 shows the distribution of standard scores on the exams. As would be expected, the low score was 20, the high was the 80 maximum and the mean was close to 50 (49.91).

Slide 12 shows the distribution of performance. The low point was 17.75, the high point was the 92 maximum, with a mean of 66.66.

Slide 13 shows the distribution of experience which is the sum of all the experience components after each was transformed by an equation. As would be expected, the distribution is quite skewed with a low point of 43.00, a high of 81.50, and a mean of 55.58. In this sample no one attained the maximum possible of 92 points for experience.

Slide 14 shows the distribution of FMS's; with a low of 107.08, a high of 232.01, and a mean of 172.15. Although the theoretical minimum of 99 was almost reached, the maximum of 264 was not approached.

Slide 15 was constructed to try to display some of the many answers possible to the question of how well we had met the desired percentages of 30, 35, and 35 for exam, performance, and experience at the E6 level.

NETPDC SEA PT STUDY--SERIES 69-FEB 75-E6 MALES N=16329 (98.54% OF E6) VB. 15 EXAM FACTOR SS 15

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(60.000)
16329	49.9067	9.9351	98.7059	0.2007	-0.1525	20.000 TO 80.000

LOWEST EXACT LIMIT = 19.50000 INTERVAL = 3.00000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
20.000	22.000	11	11	0.067	0.067	X
23.000	25.000	30	41	0.184	0.251	X
26.000	28.000	99	140	0.606	0.857	XXX
29.000	31.000	250	390	1.531	2.388	XXXXX
32.000	34.000	487	877	2.982	5.371	XXXXXXXXXXX
35.000	37.000	840	1717	5.144	10.515	XXXXXXXXXXXXXXXXX
38.000	40.000	1220	2937	7.471	17.986	XXXXXXXXXXXXXXXXXXXXXXXXX
41.000	43.000	1467	4404	8.984	26.970	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
44.000	46.000	1786	6190	10.938	37.908	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
47.000	49.000	1977	8167	12.107	50.015	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
50.000	52.000	1867	10034	11.434	61.449	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
53.000	55.000	1669	11703	10.221	71.670	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
56.000	58.000	1449	13152	8.874	80.544	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
59.000	61.000	1081	14233	6.620	87.164	XXXXXXXXXXXXXXXXXXXXX
62.000	64.000	792	15025	4.850	92.014	XXXXXXXXXXXXXXXX
65.000	67.000	573	15598	3.509	95.523	XXXXXXXXXXX
68.000	70.000	353	15951	2.162	97.685	XXXXXXX
71.000	73.000	179	16130	1.096	98.781	XXXX
74.000	76.000	110	16240	0.674	99.455	XXX
77.000	79.000	50	16290	0.306	99.761	X
80.000	82.000	39	16329	0.239	100.000	X

25

SLIDE 11

NETPDC SEA PT STUDY--SERIES 69-FEB 75-E6 MALES N=16329 (98.54% OF E6) VB. 16 PERF FACTOR 16

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(74.250)
16329	66.6634	9.5988	92.1364	-0.1943	0.2003	17.750 TO 92.000

LOWEST EXACT LIMIT = -0.00500 INTERVAL = 3.00000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

	L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
1	15.000	17.990	1	1	0.006	0.006	X
	18.000	20.990	4	5	0.024	0.031	X
	21.000	23.990	8	13	0.049	0.080	X
	24.000	26.990	3	16	0.018	0.098	X
	27.000	29.990	1	17	0.006	0.104	X
	30.000	32.990	5	22	0.031	0.135	X
	33.000	35.990	3	25	0.018	0.153	X
	36.000	38.990	27	52	0.165	0.318	X
	39.000	41.990	58	110	0.355	0.674	X
	42.000	44.990	124	234	0.759	1.433	XXX
	45.000	47.990	221	455	1.353	2.786	XXXXX
	48.000	50.990	381	836	2.333	5.120	XXXXXXXXX
	51.000	53.990	676	1512	4.140	9.260	XXXXXXXXXXXXXXXX
	54.000	56.990	1010	2522	6.185	15.445	XXXXXXXXXXXXXXXXXXXXXXXXX
	57.000	59.990	1388	3910	8.500	23.945	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	60.000	62.990	1770	5680	10.840	34.785	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	63.000	65.990	1945	7625	11.911	46.696	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	66.000	68.990	2044	9669	12.518	59.214	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	69.000	71.990	1875	11544	11.483	70.696	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	72.000	74.990	1568	13112	9.603	80.299	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	75.000	77.990	1248	14360	7.643	87.942	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	78.000	80.990	868	15228	5.316	93.257	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
	81.000	83.990	578	15806	3.540	96.797	XXXXXXXXXXXXXX
	84.000	86.990	329	16135	2.015	98.812	XXXXXXX
	87.000	89.990	146	16281	0.894	99.706	XXX
	90.000	92.990	48	16329	0.294	100.000	X

26

46

45

SLIDE 12

NETPOC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6) VB. 17 XPER FACTOR 17

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(36.500)
16329	55.5830	6.8806	47.3425	0.7327	-0.2484	43.000 TO 81.500

LOWEST EXACT LIMIT = 41.49500 INTERVAL = 1.00000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
42.500	43.490	1	1	0.006	0.006	X
43.500	44.490	21	22	0.129	0.135	X
44.500	45.490	114	136	0.698	0.833	XXXXX
45.500	46.490	217	353	1.329	2.162	XXXXXXXXX
46.500	47.490	651	1004	3.987	6.149	XXXXXXXXXXXXXXXXXXXXXXX
47.500	48.490	1262	2266	7.729	13.877	XXXXXXXXXXXXXXXXXXXX

NETPDC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6) VB. 7 FMS 07

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(124.930)
16329	172.1532	16.3235	266.4568	0.0222	-0.0970	107.080 TO 232.010

LOWEST EXACT LIMIT = 67.49500 INTERVAL = 5.00000

NUMBER OF CASES: WITHIN RANGE - 16329
OUTSIDE RANGE - 0

(X = 1750TH MODAL FREQ)

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	
102.500	107.490	1	1	0.006	0.006	X
107.500	112.490	1	2	0.006	0.012	X
112.500	117.490	7	9	0.043	0.055	X
117.500	122.490	11	20	0.067	0.122	X
122.500	127.490	32	52	0.196	0.318	X
127.500	132.490	50	102	0.306	0.625	X
132.500	137.490	148	250	0.906	1.531	XXXX
137.500	142.490	268	518	1.641	3.172	XXXXXX
142.500	147.490	532	1050	3.258	6.430	XXXXXXXXXXXX
147.500	152.490	811	1861	4.967	11.397	XXXXXXXXXXXXXXXXXX
152.500	157.490	1155	3016	7.073	18.470	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
157.500	162.490	1585	4601	9.707	28.177	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
162.500	167.490	1799	6400	11.017	39.194	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
167.500	172.490	2002	8402	12.260	51.454	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
172.500	177.490	1932	10334	11.832	63.286	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
177.500	182.490	1669	12003	10.221	73.507	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
182.500	187.490	1389	13392	8.506	82.014	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
187.500	192.490	1097	14489	6.718	88.732	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
192.500	197.490	779	15268	4.771	93.502	XXXXXXXXXXXXXXXXX
197.500	202.490	531	15799	3.252	96.754	XXXXXXXXXXXXX
202.500	207.490	300	16099	1.837	98.591	XXXXXX
207.500	212.490	131	16230	0.802	99.394	XXX
212.500	217.490	70	16300	0.429	99.822	XX
217.500	222.490	19	16319	0.116	99.939	X
222.500	227.490	5	16324	0.031	99.969	X
227.500	232.490	5	16329	0.031	100.000	X

28

50

SLIDE 14

49

COMPARISON OF DESIRED FACTOR PERCENT WITH VARIOUS STATISTICS
PAYGRADE E6 - Male Petty Officer First Class - Feb 75 - N = 16,329

FACTOR DESIRED FACTOR PERCENT	EXAM 30.0	PERF 35.0	XPER 35.0
FACPCT = 100*FACTOR/FMS			
Median (50th Percentile)	28.8	38.7	32.0
Mean	28.9	38.7	32.4
Middle 50% (25/75 Percentiles)	25.8/31.9	35.9/41.5	29.6/34.7
Low - High	12.3-44.1	13.3-53.9	21.7-58.9
E/P/X Factor-percent Combination			
Median	28	44	28
Mode	30	35-39	35-31
Middle 50% (25/75 Percentiles)	25/31	45/43	30/26
Multiple Regression			
$\Lambda\beta$ (Criterion FMS)	.410	.395	.195
$\Lambda\beta$ (Criterion Adv/not Adv)	.072	.035	.004
$\Lambda\beta/R^2$ (Criterion Adv/not Adv: R^2 = .111)	.648	.314	.037

29

SLIDE 15

Whenever policy makers set a-priori, or desired, weights there is always the question of what do they really think these weights mean. In a discussion with the Chief of Naval Personnel after he had set these weights, he stated he thought he meant that either the median value of each factor would be the desired proportion or possibly the middle 50 percent of each distribution would have approximately that proportion of their exams, performance, and experience counting toward their FMS. Although his statement led us to the development of a maximum-point weighting model, the question of which statistic best measures the goodness-of-fit remains.

First, let us look at the results in Slide 15 of using this same E6 data and calculating for each point the factor percent represented (dividing each candidate's exam standard score by his FMS and multiplying by 100, and then his performance score and his experience score treated in the same way). As may be noted in Slides 16, 17, and 18, respectively, the median values for Exam, Performance, and Experience factor-percents were 28.8, 38.7, and 32.0 compared to the desired 30, 35, and 35. Similarly, the means of 28.9, 38.7, and 32.4 or the middle 50% of 25.8/31.9, 35.9/41.5 and 29.6/34.7 can be compared. From these statistics it could be interpreted that the fit was quite good although the performance was being weighted a little heavy.

Second, let us look at the results of combining each candidates three factor percents (rounded to two digits) into a single data point. Each individual had a certain percent counting for the exam, a certain percent for performance, and a certain percent for experience. For example, if

NETPDC SEA PT STUDY--SERIES 69-FEB 75-E6 MALES N=16329 (98.54% OF E6T VB. 61 EXAM FACPCT , 61

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(	31.805)
16329	28.8820	4.4142	19.4851	-0.0360	-0.0762	12.272 TO	44.077

LOWEST EXACT LIMIT = 11.45000 INTERVAL = 1.00000 NUMBER OF CASES: WITHIN RANGE - 16329
OUTSIDE RANGE - 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
11.500	12.400	1	1	0.006	0.006	X
12.500	13.400	0	1	0.000	0.006	
13.500	14.400	8	9	0.049	0.055	X
14.500	15.400	11	20	0.067	0.122	X
15.500	16.400	20	40	0.122	0.245	X
16.500	17.400	43	83	0.263	0.508	X
17.500	18.400	64	147	0.392	0.900	XX
18.500	19.400	131	278	0.802	1.702	XXXX
19.500	20.400	213	491	1.304	3.007	XXXXXXX
20.500	21.400	334	825	2.045	5.052	XXXXXXXXXX
21.500	22.400	426	1251	2.609	7.661	XXXXXXXXXXXXX
22.500	23.400	588	1839	3.601	11.262	XXXXXXXXXXXXXXXXXXX
23.500	24.400	772	2611	4.728	15.990	XXXXXXXXXXXXXXXXXXXXXX
24.500	25.400	1024	3635	6.271	22.261	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
25.500	26.400	1176	4811	7.202	29.463	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
26.500	27.400	1299	6110	7.955	37.418	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
27.500	28.400	1457	7567	8.923	46.341	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
28.500	29.400	1496	9063	9.162	55.502	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
29.500	30.400	1406	10469	8.610	64.113	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
30.500	31.400	1296	11765	7.937	72.050	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
31.500	32.400	1163	12928	7.122	79.172	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
32.500	33.400	978	13906	5.989	85.161	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
33.500	34.400	758	14664	4.642	89.803	XXXXXXXXXXXXXXXXXXXXX
34.500	35.400	579	15243	3.546	93.349	XXXXXXXXXXXXXXXX
35.500	36.400	397	15640	2.431	95.781	XXXXXXXXXXX
36.500	37.400	296	15936	1.813	97.593	XXXXXXXXX
37.500	38.400	177	16113	1.084	98.677	XXXXX
38.500	39.400	99	16212	0.606	99.283	XXX
39.500	40.400	57	16269	0.349	99.633	XX
40.500	41.400	39	16308	0.239	99.871	X
41.500	42.400	13	16321	0.080	99.951	X
42.500	43.400	4	16325	0.024	99.975	X
43.500	44.400	4	16329	0.024	100.000	X

SLIDE 16

NETPDC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6) VB. 62 PERF FACPCT 62

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(40.639)
16329	38.7042	4.1498	17.2213	-0.1766	0.8781	13.263 TO 53.902

LOWEST EXACT LIMIT = 12.45000 INTERVAL = 1.00000 NUMBER OF CASES: WITHIN RANGE = 16329
OUTSIDE RANGE = 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
12.500	13.400	1	1	0.006	0.006	X
13.500	14.400	0	1	0.000	0.006	
14.500	15.400	0	1	0.000	0.006	
15.500	16.400	1	2	0.006	0.012	X
16.500	17.400	1	3	0.006	0.018	X
17.500	18.400	5	8	0.031	0.049	X
18.500	19.400	5	13	0.031	0.080	X
19.500	20.400	2	15	0.012	0.092	X
20.500	21.400	3	18	0.018	0.110	X
21.500	22.400	0	18	0.000	0.110	
22.500	23.400	1	19	0.006	0.116	X
23.500	24.400	7	26	0.043	0.159	X
24.500	25.400	12	38	0.073	0.233	X
25.500	26.400	21	59	0.129	0.361	X
26.500	27.400	36	95	0.220	0.582	X
27.500	28.400	49	144	0.300	0.882	XX
28.500	29.400	94	238	0.576	1.458	XXX
29.500	30.400	165	403	1.010	2.468	XXXXX
30.500	31.400	237	640	1.451	3.919	XXXXXXX
31.500	32.400	413	1053	2.529	6.449	XXXXXXXXXXXX
32.500	33.400	614	1667	3.760	10.209	XXXXXXXXXXXXXXXXXX
33.500	34.400	764	2431	4.679	14.888	XXXXXXXXXXXXXXXXXXXX
34.500	35.400	1067	3498	6.534	21.422	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
35.500	36.400	1338	4836	8.194	29.616	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
36.500	37.400	1420	6256	8.696	38.312	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
37.500	38.400	1566	7822	9.590	47.903	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
38.500	39.400	1564	9386	9.578	57.481	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
39.500	40.400	1476	10862	9.039	66.520	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
40.500	41.400	1347	12209	8.249	74.769	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
41.500	42.400	1133	13342	6.939	81.707	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
42.500	43.400	969	14311	5.934	87.642	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
43.500	44.400	741	15052	4.538	92.180	XXXXXXXXXXXXXXXXXXXX
44.500	45.400	495	15547	3.031	95.211	XXXXXXXXXXXXX
45.500	46.400	334	15881	2.045	97.256	XXXXXXXXXX
46.500	47.400	231	16112	1.415	98.671	XXXXXX
47.500	48.400	98	16210	0.600	99.271	XXX
48.500	49.400	65	16275	0.398	99.669	XX
49.500	50.400	38	16313	0.233	99.902	X
50.500	51.400	8	16321	0.049	99.951	X
51.500	52.400	7	16328	0.043	99.994	X
52.500	53.400	0	16328	0.000	99.994	
53.500	54.400	1	16329	0.006	100.000	X

32

56

57

SLIDE 17

NETPDC SEA PT STUDY--SERIES 69-FEB '75-E6 MALES N=16329 (98.54% OF E6T VB. 63 XPER FACPCT 63

N	MEAN	ST DEV	VARIANCE	SKEWNESS	KURTOSIS	RANGE(37.206)
16329	32.4138	3.8083	14.5032	0.6513	0.9492	21.685 TO 58.891

LOWEST EXACT LIMIT = 21.45000 INTERVAL = 1.00000

NUMBER OF CASES: WITHIN RANGE - 16329
OUTSIDE RANGE - 0

L SCR	U SCR	FREQ	CUM FQ	PCT	CUM PCT	(X = 1/50TH MODAL FREQ)
21.500	22.400	4	4	0.024	0.024	X
22.500	23.400	18	22	0.110	0.135	X
23.500	24.400	58	80	0.355	0.490	XX
24.500	25.400	153	233	0.937	1.427	XXXX
25.500	26.400	347	580	2.125	3.552	XXXXXXXXX
26.500	27.400	653	1233	3.999	7.551	XXXXXXXXXXXXXXXX
27.500	28.400	1055	2288	6.461	14.012	XXXXXXXXXXXXXXXXXXXXXXXXXXX
28.500	29.400	1493	3781	9.143	23.155	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
29.500	30.400	1709	5490	10.466	33.621	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
30.500	31.400	1736	7226	10.631	44.253	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
31.500	32.400	1796	9022	10.999	55.251	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
32.500	33.400	1537	10559	9.413	64.664	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
33.500	34.400	1414	11973	8.659	73.324	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
34.500	35.400	1177	13150	7.208	80.532	XXXXXXXXXXXXXXXXXXXXXXXXXXXXX
35.500	36.400	866	14016	5.303	85.835	XXXXXXXXXXXXXXXXXXX
36.500	37.400	691	14707	4.232	90.067	XXXXXXXXXXXXXXXX
37.500	38.400	518	15225	3.172	93.239	XXXXXXXXXXXX
38.500	39.400	364	15589	2.229	95.468	XXXXXXXXX
39.500	40.400	268	15857	1.641	97.109	XXXXXX
40.500	41.400	169	16026	1.035	98.144	XXXX
41.500	42.400	110	16136	0.674	98.818	XXX
42.500	43.400	77	16213	0.472	99.290	XX
43.500	44.400	46	16259	0.282	99.571	X
44.500	45.400	25	16284	0.153	99.724	X
45.500	46.400	13	16297	0.080	99.804	X
46.500	47.400	9	16306	0.055	99.859	X
47.500	48.400	8	16314	0.049	99.908	X
48.500	49.400	2	16316	0.012	99.920	X
49.500	50.400	3	16319	0.018	99.939	X
50.500	51.400	1	16320	0.006	99.945	X
51.500	52.400	3	16323	0.018	99.963	X
52.500	53.400	2	16325	0.012	99.975	X
53.500	54.400	0	16325	0.000	99.975	
54.500	55.400	1	16326	0.006	99.982	X
55.500	56.400	2	16328	0.012	99.994	X
56.500	57.400	0	16328	0.000	99.994	
57.500	58.400	0	16328	0.000	99.994	
58.500	59.400	1	16329	0.006	100.000	X

33

SLIDE 18

someone had 29% for his exam, 44% for performance, and 27% for experience these percents total 100%, and a data point of 294427 would be used for this individual in a distribution of E/P/X Factor-percent Combinations. Using an interval of five for the performance digits gave the results indicated in Slide 15 of a median combination value of 28, 44, and 28. The modal value was 30 for exam, the 35-39 interval for performance, and the experience forced to a 35-31 interval. Thus, the individuals at the lowest end of this interval had a 30, 35, 35 combination (exactly the desired percents) and the individuals at the highest end of this interval had a 30-39-31 combination. Again, as in the first method, these statistics indicate quite a good fit with performance being a bit too heavy.

And finally, looking at a third method using Multiple Regression in which $r\beta$ (the r is the correlation between a given variable and the criterion, and the β, for Beta weight, is the weight for that variable in a regression equation to predict the criterion) represents the proportion of variance accounted for by a given variable. When the criterion was the FMS in this sample the $r\beta$ proportions were .410, .395, and .195, which indicates the exam and performance were overweighted and experience underweighted. When the criterion was Advance/Not Advance the multiple R correlation was .334 and the $r\beta$ proportions were .072, .039, and .004 accounting for only 11% of the total variance. Dividing each of the above proportions by the multiple-R-squared gives .648, .314, and .039, which could be interpreted to mean that the exam was highly weighted and that experience was receiving a negligible weight.

Let me say in conclusion, that the Chief of Naval Personnel considered

the first two methods adequately demonstrated that we had given him what he had wanted when he set the a-priori weights. And perhaps, even more important, the Navy considers the advancement system to be a fair and visible system which can be understood and communicated to the sailors taking the advancement examinations how much each factor counts toward his advancement opportunity.

WHAT IS THE VALUE OF APTITUDE TESTS?

Raymond E. Christal
Occupation and Manpower Research Division
Air Force Human Resources Laboratory
Lackland AFB, Texas

INTRODUCTION

The title of my paper is "What is the Value of Aptitude Tests?" No one could feel comfortable dealing with such a broad and controversial topic--especially in front of a group of professionals in the testing business--but I feel the topic needs to be discussed and debated.

Recently, some individuals have gone so far as to suggest that testing be done away with altogether. Good heavens! Haven't we demonstrated for decades the value of tests in personnel selection and classification? Of course we must deal with reasonable questions concerning the fairness and job relevance of tests, but surely all military managers should see that tests are indispensable.

Evidently, we have done an inadequate job in merchandising our product. For this reason, I would like to look at the manner in which we have attempted to sell the value of tests and see if there are holes in our case. Then, I will venture to make a few suggestions for reorientation of our sales pitch and research strategies.

PRESENT DEFENSE

As I review the situation, I find that we have defended the value of aptitude tests on three grounds: (1) their ability to predict performance on the job; (2) their ability to predict attrition in training; and (3) their ability to predict course grades. I would like to consider these one at a time.

Prediction of Job Performance

First, let's consider job performance. Now let's be honest about it. We really don't have overpowering evidence that our tests predict job performance, and informed managers and operators know that we don't. Many of these individuals are of the opinion that the key to productivity is not individual differences in aptitude, but good management. Experience teaches them that nearly all personnel they deal with on a day-by-day basis could get the job done if they simply applied themselves. The individual differences they observe are mostly motivational, or else are not job related.

Of course, these managers are right. What they fail to understand is that this lack of variance is, to a large extent, the product of testing and training. If managers in an electronics maintenance occupation were to receive a random sample of untrained personnel out of the general population and attempt to generate the required skills on the job, I can assure you that they would quickly become acutely aware of individual differences in aptitude. However, this would not be an efficient way to run a military service. We use tests to select and classify individuals into occupations such that each person has the capacity to acquire the necessary skills for acceptable job performance. The training program, in turn, is geared to provide each trainee with these required skills. If the process is efficient, then there is no reason why tests should predict performance variance on the job, and we should neither make apologies nor hang our heads in shame when such is found to be the case.

Prediction of Attrition

The second way we have defended our tests is by showing how well they predict attrition in training. In the Air Force, a washout in pilot training costs the service thousands of dollars, and the claim is made that millions of dollars of additional costs are avoided each year by using tests to screen out applicants likely to fail in training. On the surface, this sounds like a strong case for tests. It can be shown that within any training class, individuals with high aptitude scores wash out at a much lower rate than individuals with low scores. It is also true that washouts are very expensive. However, it is not easy to demonstrate that our aptitude tests save money by reducing washout rates.

Let me show you some data extracted from the Army Air Forces Aviation Psychology Program Research Report No. 2 (DuBois, 1947).

Table 1. Attrition Rates and Aptitude Input for Every Third Pilot Training Class (44C Thru 45G)*

Class	N	Aptitude Cutoff	Percent Eliminees
44C	12,232	3	15.5
44F	9,371	3	12.0
44I	6,466	4	19.6
45A	6,525	4	21.0
45D	1,384	4	21.5
45G	664	6	27.4

*Extracted from Report No. 2, "The Classification Program," Army Air Forces Aviation Psychology Program Research Reports, 1947.

Table 1 reflects pass/fail data for every third class from 44C through 45G. In classes 44C and 44F, the cutting score on the aptitude score for entry was Stanine 3, and the average attrition rate was 13.9%. In classes 44I, 45A, and 45D, the cutoff was raised to Stanine 4. However, instead of going down, the attrition rate increased to 20.4%. Finally, in class 45G, the cutting point was raised to Stanine 6, yet the attrition rate went up again--clear up to 27.4%. In view of these data, one might conclude that attrition in pilot training would be minimized if those cases having the least aptitude were entered into training.

Of course, this is not true. The fact is that attrition rates were controlled by administrative actions, and were not dependent on the quality of the input. The number of pilot graduates was determined in large part by the number of cockpits to be filled. The data shown in Table 1 reflect actions taken toward the end of the war as the number of trained pilots became abundant and aircraft production was reduced. We have good reason for believing that the quality of graduates from these classes varied, but we cannot demonstrate that the use of tests saved money by reducing attrition rates.

We would have even a more difficult time demonstrating the influence of tests on attrition rates in enlisted courses. The number of graduates from such courses is ordinarily programmed months in advance to meet operational requirements, and fluctuations in input talent produce only minor fluctuations in attrition rates. During periods of low quality input, it is not uncommon to increase wash-backs and remedial training to maintain production standards.

Pass/fail is a very slippery criterion, and attrition rates seem to be arbitrarily established. This phenomenon is not restricted to the military. For example, there are wide variations in the input talent to colleges and universities, where attrition rates for the same courses are essentially equivalent. A washout from MIT or Cal Tech could be an honor graduate from certain other colleges and universities. We seem to be living in a relative world without absolute standards. This is one of the problems we face in demonstrating the value of tests.

In 1957, Dr. Krumboltz and I published a study (Krumboltz and Christal, 1957) in which we demonstrated that the probability of a student completing pilot training is a function of the aptitude levels of the other three students with whom he is grouped under the same instructor. A student with a Stanine 5 was less likely to graduate if he were grouped with three students at the Stane 9 level than if he were grouped with three students at the Stanine 5 level.

In 1959, an investigator in Australia reported a strange and related finding (Want, 1959). In that country, Air Force and Navy pilots were being trained together under the same instructors. The Air Force raised their entrance requirements, and the result was that the attrition rate for Navy trainees nearly doubled. While the level of talent of Navy trainees remained constant, these individuals began looking bad in comparison with their Air Force counterparts.

These studies demonstrate that aptitude tests do measure differences in abilities which are recognized by instructors. However, we will not be able to defend our tests on the basis of their role in reducing attrition rates until absolute standards for successful course completion are implemented and adhered to.

Prediction of Course Grades

A third way to have attempted to show the value of tests is in terms of their ability to predict final course grades. The statement that aptitude tests predict course grades is irrefutable. Literally hundreds of studies have consistently demonstrated this to be so. To prove that we haven't lost our grip in this respect, I've brought along results from one of the largest Air Force validation studies ever conducted, which I will display to you.

We began with a 380,000-case population graduating from Air Force entry-level courses between January 1969 and April 1974. From this population, we randomly selected 1,000 cases from each course, when available, or a total sample when data were available from fewer than 1,000 cases. This yielded a total validation sample of slightly more than 100,000 cases, representing graduates from 134 different courses.

Table 2. Validities (R) of AQE/ASVAB/AFQT for Course Grades* for AFSs with AQE/ASVAB Cutoff of 80th Centile

R	N	R	N
.626	59	.422	1000
.543	303	.422	1000
.507	679	.421	1000
.485	749	.414	1000
.483	168	.409	1000
.483	426	.407	988
.472	1000	.406	1000
.471	1000	.396	152
.471	249	.395	503
.471	1000	.394	217
.463	1000	.387	1000
.457	509	.386	1000
.456	1000	.383	1000
.444	1000	.382	753
.439	624	.379	637
.438	1000	.374	1000
.437	1000	.366	209
.435	1000	.348	716
.431	1000	.324	1000
.429	608	.285	283
.427	777	.164	1000

Median R = .424 Total N = 31,609

*For cases graduating between Jan 1969 and Apr 1974.

The validity coefficients I will show are uncorrected multiple correlation coefficients for a weighted composite of the four AQE composites and AFQT against final course grades. The values in Table 2 show the validities computed in 42 courses for which the cutting score on AQE was at the 80th centile. These coefficients may look a little low, but remember that they are uncorrected and have been computed in a sample which has been subjected to severe restriction in range on the predictors. Since the bivariate normality assumptions could not be met, no corrections for restriction were made. However, it is estimated that in an unrestricted population, many of these validities would be found to be in the .60s, .70s, and .80s. The median correlation obtained in the computing sample was .42. The lowest reported validity is for a Linguistic/Interrogator course for which the Air Force has special additional screening procedures.

Table 3. Validities (R) of AQE/ASVAB/AFQT for Course Grades* for AFSs with AQE/ASVAB Cutoff of 60th or 70th Centile

R	N	R	N	
.647	78	.439	210	
.631	139	.439	1000	
.624	658	.435	1000	
.619	163	.422	697	
.586	434	.415	823	
.551	1000	.410	146	
.535	605	.405	129	
.531	1000	.392	412	
.529	606	.389	1000	
.527	908	.386	999	
.527	332	.385	1000	
.518	1000	.381	425	
.518	1000	.370	114	
.517	1000	.348	1000	
.502	892	.327	1000	
.498	612	.305	1000	
.492	1000	.305	1000	Comp Operator
.491	65	.232	228	Comp Programmer
.484	539	.176	202	Small Arms
.474	1000	.173	1000	AC&W Operator
.458	291	.158	1000	Radio Operator
.440	1000			

Median R = .440 Total N = 28,707

*For cases graduating between Jan 1969 and Apr 1974.

Table 3 reports uncorrected validities for 36 courses having entry-level requirements at the 60th or 70th centile on AQE. Again, these coefficients are attenuated by severe restrictions in range, although some of the uncorrected Rs are higher than .50.

I might point out that five of the lowest six coefficients in this table are associated with courses training students in operator-type jobs. Two are for radio and morse system operators, for which a special code test is available to enhance prediction of student success. The other three are for computer operators, aircraft control and warning operators, and small arms specialists. In each instance, certain perceptual-psychomotor skills are required which are not measured by the AQE or AFQT.

The median uncorrected validity of the tests for these 42 schools was .44, which, again, is a gross underestimate of values which would have been obtained in an unrestricted sample.

Table 4. Validities (R) of AQE/ASVAB/AFQT for Course Grades* for AFSs with AQE/ASVAB Cutoff of 40th or 50th Centile

R	N	R	N	R	N
.678	807	.557	532	.488	1000
.672	636	.556	1000	.482	628
.668	105	.552	437	.479	1000
.657	100	.550	177	.465	850
.652	140	.549	641	.465	1000
.634	980	.544	1000	.440	1000
.628	1000	.542	1000	.432	305
.625	649	.536	666	.432	814
.592	1000	.535	240	.422	1000
.591	1000	.532	1000	.412	1000
.588	532	.531	1000	.404	890
.584	886	.528	598	.392	1000
.581	1000	.527	376	.378	609
.574	1000	.521	1000	.375	1000
.572	1000	.498	208	.371	1000
.570	715	.493	563	.369	191
.568	1000	.490	1000	.351	372
.566	1000	.489	751	.263	1000
.565	575	.489	1000	.221	1000

Median R = .532 Total N = 42,973

*For cases graduating between Jan 1969 and Apr 1974.

Table 4 reports validities for grades in 56 courses for which AQE entrance requirements are at the 40th or 50th centile levels. These coefficients are higher because they are less subject to restriction in range. The median value is .53. However, these coefficients are considerably below what would be obtained in an unrestricted sample. Not only have the lower 40 to 50 percent of the standardization population been denied entry into the course, but the number of cases in the upper levels of the aptitude distribution is severely limited due to siphoning off by more demanding courses.

Once again, by the data I have presented, we can demonstrate that aptitude scores predict course grades. I'm not sure, however, that this fact impresses the average military manager. After all, one cannot translate course grade points into dollars and cents or manpower bodies; nor have we been able to demonstrate convincingly that graduates with high course grades actually perform better on the job than graduates with low course grades, even though they in fact may do so.

Summary of Current Status

So here we stand. Although we feel that aptitude tests predict job performance, we have very little data to support this contention. We would like to claim that the use of tests reduces attrition in training, but the evidence suggests that attrition rates are primarily a function of administrative actions, not level of input talent. We can show that test scores predict course grades, but this doesn't seem to impress the average military manager. Where do we go from here?

SUGGESTED CRITERIA FOR TEST EVALUATION

It would be my recommendation that, in the future, we focus our attention on five types of criteria for test evaluation as follows:

1. Speed of skill acquisition
2. Speed of skill decay — Skills Maintenance
3. Speed of skill reacquisition — Skills Maintenance
4. Speed of response — Performance
5. Accuracy of response — Performance

Speed and accuracy of response may be important in some occupations involving a demand for perceptual-psychomotor or clerical skills. However, due to time limitations, I have elected to address only the first three criteria, which relate to the speed of skill acquisition, decay, and reacquisition. In all three instances, the basic variable against which tests are to be evaluated is TIME. Time is an excellent criterion. It has a zero point; it can be measured in equal intervals; it is easily understood by military managers; it can be easily converted into dollars and cents or manpower spaces; and it is the single most expensive item in the military budget.

The military services spend literally billions of dollars each year supporting the development and maintenance of skills. The more obvious expenditures are associated with formal residence and on-the-job training courses, but this is just the top of the iceberg. For example, the Air Force spends hundreds of millions of dollars each year just to maintain pilot and navigator skills. Even more costly is the time individuals in all services spend in learning to perform new tasks as they are encountered on a day-by-day and assignment-by-assignment basis. To the extent that aptitude scores predict the time required for individuals to acquire and maintain skills, they can be used to reduce costs and optimally distribute talent to jobs. I will address this issue during my remaining time.

Skills Acquisition

There is nothing unique or new about the concept of aptitude scores predicting learning rate. For example, in 1963, John B. Carroll recommended that aptitude be defined as learning rate (Carroll, 1963). The first intelligence test developed by Alfred Binet, back in 1904, was designed to measure differences in the level of skills acquired by individuals during a constant time interval (chronological age). These scores were later normed and converted into a score "mental age." A ratio of the mental age to chronological age was computed and came to be called the Intelligence Quotient (IQ). Regardless of the problems associated with the development and utilization of IQ scores, they have been used for years as rough indicators of individual learning rates.

In the academic world, many tests are called learning abilities measures, and have been used for decades by teachers to place pupils into homogeneous groups so as to minimize variance in learning rates within groups. Tests have been shown to be valid predictors of school grades, both in the academic world of the civilian sector and in all military services, and school grades can be viewed as the amount of content mastered by students when learning time is held constant. Aptitude tests also predict proficiency test scores in the services, which are rough measures of the amount of content mastered by individuals at various career points. In Project UTILITY (Vineberg and Taylor, 1972), which was conducted for the U.S. Army by the Human Resources Research Organization in the late 1960s, AFQT scores were shown to be related to the rate of skill acquisition in several occupational areas. However, with the passage of time, an increasing proportion of men at all levels of AFQT appeared in the upper ranges of performance distributions, indicating that for these low-level occupations, aptitude scores predict the rate of skills acquisition, but not ultimate level of performance. Pilot training programs are generally locked-step. For this reason, I have been unable to locate data demonstrating that aptitude scores predict speed of skill acquisition. However, pilot aptitude tests do predict within-class elimination for flying deficiency, and individuals in the flying research area assure me that slowness in acquiring skills is the primary cause for such elimination. This observation needs to be confirmed by carefully controlled research.

While the evidence that aptitude scores predict learning time is substantial, most of it is indirect. Outside of a few laboratory experiments dealing with paired associates learning, I have been able to locate few studies directly addressing the subject, and these have involved small Ns and produced mixed results. In one study conducted by a graduate student at the University of Pittsburg

(Wang, 1968) and in another study conducted by the Human Resources Research Organization (Wagner, Behringer, & Pattie, 1973) substantial relationships were found between general and specialized aptitude tests and learning times; however, there appeared to be complex interactions among learning rates, types of materials to be learned, training modalities, and various aptitude scores. If such findings are generally confirmed, the proper selection and classification of personnel may be more complicated than it appears on the surface. However, in one unpublished study conducted by the Navy[1], no such interactions were found, and standard Navy aptitude tests were demonstrated to have substantial validity for predicting training times (see Table 5). This study involved two tracts in a Navy aviation familiarization course, one which was made up solely of reading modules, and the second which included seven slide/tape modules.

Table 5. Validities of Aptitude Scores for Time (Hrs) Criteria in Navy Aviation Familiarization Course

Group #1 - 7 Slide/Tape + 9 Reading Modules (N=109)

Aptitude Test	Validity
GCT	-.58
Arithmetic	-.47
Clerical	-.34
Multiple R	-.67

Group #2 - 16 Reading Modules (N=113)

Aptitude Test	Validity
GCT	-.45
Arithmetic	-.43
Clerical	-.26
Multiple R	-.51

Multiple Rs and Cross-Application Rs

	Development Sample R	Cross-Application Sample R
#1	.67	.66
#2	.54	.53

[1]Information in this table were provided by Dr. Kirk A. Johnson, Navy Personnel Research and Development Center, Memphis Branch Office, Millington, Tennessee. Multiple Rs and cross-application Rs were computed by the author using the correlation matrices provided by Dr. Johnson.

Interestingly, the higher validities were obtained for the slide/tape group. Notice that the equations predicting time criteria for the two treatments were highly homogeneous.

I was also able to obtain data for a 200-case sample of Air Force personnel who recently completed an individualized instruction course (Inventory Management) at Lowry Air Force Base. Two criteria were available, one of which was a summation of time to complete the course blocks, and the other of which was a summation of course block scores (grades). The results of this analysis are presented in Table 6.

Table 6. Validities of ASVAB/AFQT Scores for Time and Grade Criteria in the Air Force Inventory Management Course (N=200)

Development Sample Rs

Criterion	Predictors	Multiple R
Time	General AI, Electronic AI, AFQT	.39
Grade	General AI, Electronic AI, Mechanical AI	.59

Cross-Application Rs

Source of Predictive Weights	Application Criterion	R
Time Criterion Predictors	Grade	.37
Grade Criterion Predictors	Time	.55

The multiple validity of the ASVAB composites and AFQT for the training time criterion was only .39--which was significant, but lower than hoped for. However, the multiple validity of three ASVAB composites for the sum of block test grades was .59, which is higher than was obtained for final school grades when the course was taught in a locked-step fashion. Even though this course is now taught in an individualized instruction mode, there appears to be more predictable variance in the amount of content mastered than in the time for course completion. This finding is explained, in part, by the fact that students in the course took module and block tests when they felt they were ready for examination. Upon first testing, some students barely reached a 70% passing standard, while others routinely scored 100% on many tests. These latter students had reached the 70% standard at much earlier (but unknown)

points in time, so there was no simple way to compute a time-to-standard for each case, In this sample, the correlation between the time and grade criteria was -.40, indicating that students completing the course in the shortest time tended to be those who mastered the greatest amount of content.

There is not time to discuss problems associated with generating a pure time-to-standard criterion in the operational setting, but I would like to recognize that such problems do exist. It is unlikely that individualized instruction courses presently train all students to exactly the same standard (although some meet a 90-90 standard), even though finishing times may vary. Until this problem is resolved, it will be difficult to establish the exact relationship between aptitude scores and learning rates in such courses. Ultimate solutions may include better records and controls, continuous testing, statistical corrections, and controlled experiments. One must admit that the problems to be overcome are challenging.

It should be observed from Table 6 that the equations predicting the grade and time criteria are homogeneous. This provides additional evidence that, since tests normally have high validity for course grades, they should also be found to be highly related to learning time criteria. It is important, however, that direct relationships be established. The author would appreciate receiving copies of any studies bearing on the question.

Prediction of Decay Rates

A second stream of research which needs to be initiated concerns the ability of aptitude tests to predict decay rates for skills and knowledges. There has been a great deal of research leading to the development of generalized curves of retention, but surprisingly little research has been accomplished relating to individual differences in retention. Underwood published one summary paper (Underwood, 1954) in which he concludes that, when associative strength is held constant, there are no differences in forgetting rates as a function of aptitude during the first 24 hours. However, this study dealt with laboratory associative learning experiments and short decay periods. The military services should be able to provide more definitive answers concerning individual differences in forgetting rate as a function of aptitude.

One very revealing study was reported by the Naval Personnel and Training Research Laboratory in 1970 (Johnson, 1970) which provided data relating to the skill decay question. The study was based on material being taught in the first phase of the avionics fundamentals course. Proficiency was measured by means of the

criterion referenced tests that had been used to validate the programmed instructional material used in this phase. Measures were obtained on a pre-test, on an immediate post-test, and at intervals of 1 day, 7 days, 28 days, and 96 days following the original learning. It was found that in spite of a fairly high level of mastery on the immediate post-tests and a considerable amount of review, much of the material learned during the first phase of the course was forgotten by the end of the course. The differences between individual students were large on the pre-test, were quite small on the immediate post-test, and increased gradually over the remaining post-tests until, by the end of the course, they were almost as large as they were on the pre-test.

Although this study was based on only a fairly small N, it did provide a set of relatively unique data. The experiment began with 141 students. Seven were dropped for administrative reasons; eight failed because of slow progress; 21 washed back because of slow progress; and 17 were moved ahead because of fast progress. Thus, only 85 cases were left in the final sample, and these cases were fairly homogeneous in terms of learning rate. In spite of this homogenization process, data in the study can be re-analyzed to reflect differential decay rates as a function of aptitude. As can be seen in Table 7, aptitude scores account for 24% of the final test score variance, with original pool test scores held constant (partial multiple R^2). Although one might argue that associative strength was not held constant, from a practical standpoint it can be stated that individuals showed differential decay rates in criterion referenced test scores as a function of their aptitude levels.

Table 7. Retention of Electronics Fundamentals as a Function of Aptitude

Predictors	Validities for Final Post-Test	
	R^2	R
Immediate Post-Test	.185	.430
Aptitude Tests Alone	.312	.559
Immediate Post-Test Plus Aptitude	.382	.618
Unique Contribution of Aptitude Tests	.197	.444
Aptitude Tests with Immediate Post-Test Scores Held Constant (Partial R^2 and R)	.242	.492

Predicting Time for Reacquisition of Skills

The third area which needs to be addressed concerns the time required for reacquisition of skills and knowledges which have degenerated over time as a function of disuse. One would hypothesize that if aptitude scores predict the speed of skills acquisition, they should also predict the speed of skills reacquisition; but, to my knowledge, this has not been firmly established in the military setting. I conducted one analysis in the early 1950s which I now wish I had documented, since it bears on the question. A number of World War II pilots were recalled during the Korean conflict and sent to flight instructors' school. At the school they were given training to reestablish their flying skills. I managed to locate the original World War II pilot aptitude scores for a sample of these individuals and found, to my amazement, that they were still predictive of flying proficiency grades for students in this course--in spite of the passage of time and in spite of the original screening, training, and differential experiences these individuals had during and subsequent to World War II.

The question concerning the relationship between aptitude and the time required for skills maintenance is extremely important. For example, consider the pilot area alone, where the Air Force spends hundreds of millions of dollars per annum in terms of fuel, aircraft, and maintenance costs in order to maintain flying proficiency. In the foreseeable future, multi-millions of dollars will be spent for sophisticated simulators in hopes of saving fuel and aircraft associated with this expensive but necessary program. Yet, we know very little about the rates of skill decay and regeneration, and practically nothing concerning individual differences in such rates. Are individuals who quickly attain pilot skills also those who slowly lose such skills and quickly regain them after decay? If so, proper selection of individuals into the pilot training program may be more important than generally recognized. Because of the large numbers involved, the potential savings might be even larger on the enlisted side, although they may be more difficult to document.

SUMMARY

I realize that I have wandered far and wide in this rather loosely organized paper, but I will try to summarize briefly. I have suggested that we should begin moving away from job performance, pass/fail criteria, and school grade criteria for aptitude test evaluation. Certain types of perceptual-psychomotor tests and tests of clerical speed and accuracy may predict performance in operator and clerical type jobs; however, we should not expect tests to have predictive efficiency for performance in jobs where performance is primarily a function of the extent to which fully developed skills

are applied. Test scores do predict the relative probability of failure within training groups, but they do not determine failure rates for groups as a whole. Pass/fail rates are determined by administrative actions, rather than quality of input. Test scores predict training grades, but grade points cannot be easily translated into dollars and manpower.

I have suggested that we should demonstrate the value of tests in terms of their ability to predict personnel time requirements for skills acquisition and maintenance.

Finally, I have enumerated some of the research findings to date which bear upon critical issues, and have suggested research studies which should be undertaken.

I am personally convinced that aptitude tests are indispensable in the military setting, and that they must continue to be utilized, in spite of problems which may exist with respect to test fairness. I have faith that ways will be found to eliminate or reduce test biases which may exist. At the same time, I feel that we have an obligation to demonstrate the value of tests in terms of their ability to help us operate our military establishment in a cost-effective manner.

What is the value of aptitude tests? I cannot give a precise answer to this question; but they are of considerably more value than most military managers have been led to believe.

REFERENCES

Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64(8), 723-733.

DuBois, P. H. (editor). Army Air Forces Aviation Psychology Program Research Report No 2, Washington, D.C.: U.S. Government Printing Office, 1947.

Johnson, K. A. Retention of electronic fundamentals: differences among students, Technical Bulletin STB 71-2, Naval Personnel and Training Research Laboratory, October 1970.

Krumboltz, J. D. & Christal, R. E. Relative pilot aptitude and success in primary pilot training. Journal of Applied Psychology, 1957, 41, 409-413.

Underwood, B. J. Speed of learning and amount retained: A consideration of methodology. Psychological Bulletin, 1954, 51, 276-282.

Vineberg, R. & Taylor, E. N. Performance in four Army jobs by men at different aptitude (AFQT) levels: 3. The relationship of AFQT and job experience to job performance (HumRRO Tech. Rep. 72-22). Alexandria, Va.: Human Resources Research Organization, August 1972.

Wagner, H., Behringer, R. D., & Pattie, C. L. Individualized Course Completion Time predictions: Development of instruments and techniques (HumRRO Tech. Rep. 73-25). Alexandria, Va.: Human Resources Research Organization, November 1973.

Wang, M. L. C. C. An investigation of selected predictors for measuring and predicting rate of learning in classrooms operating under a program of individual instruction. Doctoral dissertation, International University Microfilm, 1968.

Want, R. The frames of reference of flying instructors. Journal of Applied Psychology, 1959, 43, 86-88.

STAFF NONCOMMISSIONED OFFICER TASK ANALYSIS STUDY

BY

MAJOR WILLIAM R. MASCIANGELO USMC
OFFICE OF MANPOWER UTILIZATION
MCDEC, QUANTICO, VIRGINIA 22134

The Office of Manpower Utilization (HQMC) is tasked with conducting task analysis studies in support of Manpower Planning. Our Office, since 1969, has been concerned with job structure, classification and assignment. Training has been addressed in very broad terms.

I plan, during this period, to address a task analysis study our office conducted for the Marine Corps Staff Noncommissioned Officers Academy (SNCOA) at Quantico, VA. It would be safe to assume that most everyone in this room is somewhat familiar with the methodology of conducting a task analysis, so I intend to emphasize those aspects of this study which differ from a conventional task analysis.

My presentation will follow this outline: (Slide 1)

I. Background of SNCO Task Analysis Study

II. Methodology Overview

III. Special Problems

IV. Analysis Procedures

V. Conclusions

I. Background of SNCO Task Analysis Study

In November of 1975 the Education Center at Quantico, VA requested that the Office of Manpower Utilization (MPU) conduct a task analysis study of Marine Corps SNCOs for the SNCO Academy at Quantico, VA.

The purpose of the study was (Slide 2) to identify only common managerial/supervisory and general military tasks that are being performed by Marine Corps SNCOs regardless of their Military Occupational Specialty (MOS) and to use this information to update the SNCOA curriculum.

The study was conducted by 1 officer and 1 SNCO on a part time basis. It was a pilot study in that it was the first task analysis study conducted by MPU in direct support of a Marine Corps Formal School.

Before proceeding further, let me point out the mission of the SNCOA in order that you may better appreciate the unique problems that were encountered.

> The mission of the SNCOA is (Slide 3): To provide instruction to Staff Sergeants (SSgt) and Sergeants (Sgt) selected for SSgt in the fundamental professional knowledge and skills required by the modern Marine Corps SNCO, regardless of occupational speciality.

Initially only common managerial/supervisory tasks were identified for inclusion within a task inventory. Subsequently, general military tasks pertaining to physical training, drill and marksmanship were added in order that all the tasks contained within the SNCOA Program of Instruction (POI) were reflected.

Tasks used in this study can therefore be categorized into two areas; Managerial/Supervisory and General Military Tasks. To be even more specific, further explanation of these task categories is required.

The managerial/supervisory tasks were divided into two parts; those tasks with soft verbs and those tasks with hard verbs.

Soft verbs are considered difficult to define and difficult to time rate. They are verbs used in higher level tasks as part of the management functional areas of planning, organizing, staffing, controlling and directing. Examples of these verbs used in tasks are: (Slide 4)

Hard verbs are more easily definable and there is usually common agreement about their meaning. They are also easier to time rate. Examples of these verbs used in tasks are: (Slide 5)

The General Military Tasks are tasks performed by SNCOs in the areas of physical training, drill and marksmanship. These tasks are required by HQMC directives. Examples of these tasks are: (Slide 6)

The criteria used for developing the task inventory were as follows: (Slide 7) (1) only managerial/supervisory and general military tasks were to be used, and (2) the tasks had to be performed by SNCOs from two or more occupational fields; if a task was unique to any one occupational field it was not included in the inventory.

II. Methodology Overview (Slide 8)

Task statements which met the criteria of the study were subjectively selected from previously completed task analysis studies. Managerial textbooks and POIs were also screened for possible tasks. Interviews were conducted with over 132 SNCOs from all grades (E-6 to E-9), representing 34 MOSs. The purpose of the interviews was to ensure that the tasks were understood and that the tasks met the criteria of the study. Tasks were also added and deleted during the interviews. The final task inventory booklet contained 250 task statements which were categorized into 14 duty areas. The duty areas were (Slide 9) Administration Supervisor, Administration Worker, Counselling, Training, Inspections, Physical Training, Guard/Classified Material, Briefing, Supply/Logistics, Legal, Maintenance, Parades/Drill, Embarkation and Fiscal. The inventory booklets also contained demographic, academic and job/career satisfaction questions. Booklets were administered to 1492 SNCOs (E-6 to E-9) at Camp Lejeune and Cherry Point, North Carolina. The SNCOs represented 38 of 39 occupational fields. There were 748 E-6s, 456 E-7s, 220 E-8s and 68 E-9s.

III. Special Problems (Slide 10)

There were two areas in which this study differed from a conventional task analysis. (Slide 11)

(1) It was an attempt to test the feasibility of identifying common managerial/supervisory tasks. These tasks tend to be broad and vague. Thus, the assumptions of time ratability and consistent level of specificity were severely strained.

(2) It was also an attempt to conduct a TA study in which it was not possible to account for 100% of a Marines job.

IV. Analysis Procedures (Slide 12)

The initial analysis approach was (Slide 13) to determine which tasks Staff Sergeants (E-6) and Gunnery Sergeants (E-7) performed. The SNCOA was only interested in these two grades. A duty area analysis was conducted using special job descriptions with a pay grade option. Groups are then characterized

as to the extent of performance in each duty area. A duty area analysis is just one clue as to the relative time spent in a duty area by one group compared to another group. The value of this approach is that it lets the analyst obtain a better understanding about the time a group spends in each duty area. For example, what initially might appear as a lot of time spent in a duty area, would, when compared to the study population mean time spent, actually show that the time spent was below the mean and therefore may not be as significant as initially thought.

The next slide (14) shows 7 of the 14 duty areas and the relationships between pay grades and the time spent in each duty area. Based upon what is known about the type of work performed by SNCOs in various occupational fields these findings were not surprising. The SSgt and GySgt do a little bit of everything. They are the worker supervisors. The E-8 and E-9 can be categorized as generalists since the type of tasks they performed were broad and not as specific as the E-6 and E-7. They spent their time above the population mean in only a few duty areas. They were below the population mean in the remainder of the duty areas.

The final report with the findings and recommendations was based on the analysis of the special (pay grade) job descriptions. This initial analysis was all that was required to provide the SNCOA with the information they needed to redesign their POI.

In preparation for MTA and to determine what effect violation of TA assumptions had upon the clustering process the overlap group and diagram programs were requested. (Slide 15) It had been anticipated that the SNCOs would probably group by occupational field, pay grade, and or organizational level. This was found to be true. However, the diagram, because of the TA violations showed something else.

Thirtynine stages were initially identified for further analysis. The duty area in which each of these stages spent the highest percent of time was determined. A summary of this data (Slide 16) shows that 74% of the population within the study spent the highest amount of time in the Administration Supervision duty area. This chart shows the breakdown.

	DUTY AREA	POPULATION	PERCENTAGE
1.	Admin Supervisor	1029	74
2.	Counselling	191	14
3.	Training	146	10
4.	Guard/Classified Material	28	2

The majority of the managerial/supervisory soft verb tasks are contained in these four duty areas. The significance here was that the SNCOA was not teaching managerial functions yet SNCOs were performing managerial tasks.

Closer analysis showed that the diagram could be divided into two parts. (Slide 17) Part 1 contained 659 SNCOs from a variety of occupational fields (Admin, Intel, Maint, ect). This large cluster was mostly senior SNCOs working at higher level commands. These SNCOs can be classified a generalists and performed tasks that categorized them as administrative supervisors. The tasks performed by these SNCOs were mostly in the soft verb category. The job descriptions identifying subgroups within this Part I cluster were very similiar because of the soft verb tasks. Additional programs were needed to identify the members. The tasks performed by this group represented about 85 to 90% of their job. The degree of overlap was betweem 34 and 55%.

Part II is characterized by small separate and distinct groups identifying with specific MOSs, occupational fields or billets. In this case tasks which contained hard verbs were more prevalent. Here, the tasks, even though meeting the criteria of the study, were occupational field specific and the level of specificity between these tasks and the tasks found in Part I was not the same. Some groups identified in Part II were administrators, supply chiefs, maintenance chiefs, counsellors, and instructors. These SNCOs were worker supervisors. The tasks performed by this group represented about 70% of their jobs. The degree of overlap within each group was between 15 and 29%.

To what extent did the failure to obtain 100% of the SNCOs job influence the diagram?

SNCOs were instructed to mark tasks that they performed in their present job and not what they did or thought SNCOs should do. Since the majority of SNCOs were unable to account for 100% of their job they were requested to answer, after

completion of the inventory, this question (Slide 18): The tasks you selected above represent what total percent of time in your present job? There were 233 SNCOs (Slide 19) who stated that what they marked represented 100% of their job. (Slide 19)

For information, the answer to this question ranged from an average of 68% to 88%. The population average was 77%. An occupational field comparison of this question showed that occupational field 43 was the lowest with 43% and occupational field 99 was the highest with 88%. (Slide 20) Shows a sample of the occupational field percentages.

Analyzing the occupational fields it was found that their estimates of time spent appeared to be reasonable. This is based upon what is known about the type of work each occupational field normally performs.

Even though the percent of time performing each task is inflated, the amount of time spent performing one task is relative to the time spent performing the other tasks.

V. Conclusion (Slide 21)

Although some basic assumptions about the TA methodology were violated the feasibility of addressing common managerial/supervisory tasks and not accounting for 100% of a Marines job still produced worthwhile results for the SNCOA.

If the inventory had contained tasks for all SNCOs to identify 100% of their job the diagram would have looked the same. That is, SNCOs would have grouped by occupational field but with higher degrees of overlap.

The level of specificity was raised by using managerial/supervisory tasks with soft verbs. This enabled SNCOs to cluster whose occupational fields were very different yet who still performed similiar administrative supervisory tasks.

As a result of this study the SNCOA has begun to rewrite their POI based upon what the SSgts and GySgts are actually doing.

Finally, this study will provide some point of departure for attempting similiar studies for officer schools such as Amphibious Warfare School and Command and Staff College.

PRESENTATION OUTLINE

I. BACKGROUND OF STAFF NONCOMMISSIONED OFFICER STUDY

II. METHODOLOGY OVERVIEW

III. SPECIAL PROBLEMS

IV. ANALYSIS PROCEDURES

V. CONCLUSIONS

PURPOSE

TO IDENTIFY ONLY COMMON MANAGERIAL/ SUPERVISORY AND GENERAL MILITARY TASKS THAT ARE BEING PERFORMED BY MARINE CORPS STAFF NONCOMMISSIONED OFFICERS REGARDLESS OF THEIR MILITARY OCCUPATIONAL SPECIALTY (MOS)

SNCO ACADEMY MISSION

TO PROVIDE INSTRUCTION TO STAFF SERGEANTS AND SERGEANTS SELECTED FOR STAFF SERGEANT IN THE FUNDAMENTAL PROFESSIONAL KNOWLEDGE AND SKILLS REQUIRED BY THE MODERN MARINE CORPS STAFF NCO, REGARDLESS OF OCCUPATIONAL SPECIALITY

SAMPLE TASKS

MANAGERIAL/SUPERVISORY — SOFT VERBS

1. ANALYZE SECTION WORK LOAD REQUIREMENTS
2. FORMULATE PROJECT GOALS
3. COORDINATE WORK ACTIVITIES WITH OTHER SECTIONS
4. EVALUATE PERFORMANCE OF MILITARY SUBORDINATES
5. DEVELOP WORK METHODS AND PROCEDURES

SAMPLE TASKS

MANAGERIAL/SUPERVISORY HARD VERBS

1. DRAFT REPORTS
2. PREPARE LESSON PLANS
3. ASSIGN SCHOOL QUOTAS TO SUBORDINATES
4. REQUISITION PUBLICATIONS
5. BRIEF VISITORS

SAMPLE TASKS

GENERAL MILITARY

1. SCORE PERSONNEL TAKING PHYSICAL FITNESS TEST
2. CONDUCT PLATOON DRILL
3. QUALIFY WITH WEAPON
4. VERIFY TARGET HITS/SCORES ON QUALIFICATION DAY
5. POST OR RELIEVE WATCH/DUTY STANDERS

CRITERIA FOR TASK INVENTORY CONSTRUCTION

1. ONLY MANAGERIAL/SUPERVISORY AND GENERAL MILITARY TASKS WOULD BE USED.

2. THE TASKS HAD TO BE PERFORMED BY SNCOs FROM TWO OR MORE OCCUPATIONAL FIELDS.

3. IF A TASK WAS UNIQUE TO ONLY ONE OCCUPATIONAL FIELD IT WAS NOT INCLUDED IN THE INVENTORY.

II. METHODOLOGY OVERVIEW

INVENTORY BOOKLET

250 TASKS

COUNSELLING
BRIEFING
INSPECTION
TRAINING
MAINTENANCE
GUARD/CLASSIFIED MATERIAL
ADMIN SUPERVISOR
ADMIN WORKER
PARADES AND DRILL
PHYSICAL TRAINING

14 DUTY AREAS

LEGAL
SUPPLY
EMBARKATION
FISCAL

III. SPECIAL PROBLEMS

SPECIAL PROBLEMS

1. AN ATTEMPT TO TEST THE FEASIBILITY OF IDENTIFYING COMMON MANAGERIAL/SUPERVISORY TASKS

2. AN ATTEMPT TO CONDUCT A TASK ANALYSIS STUDY IN WHICH IT WAS NOT POSSIBLE TO ACCOUNT FOR 100% OF A MARINES JOB.

IV. ANALYSIS PROCEDURES

ANALYSIS APPROACH "A"

FOR THE SNCO ACADEMY

1. DUTY ANALYSIS BASED UPON SPECIAL JOB DESCRIPTIONS (PAY GRADE OPTION)

2. ANSWER QUESTION: WHAT TASKS DO THE PAY GRADES E-6 AND E-7 PERFORM?

70

DUTY PROFILE (SAMPLE)

DUTY AREAS/% TIME PERFORMING	E-6	E-7	E-8(1STSGT)	E-9(SGTMAJ)	MEAN	S.D
COUNSELLING	13.99	13.64	19.74	27.14	14.10	2.01
BRIEFING	3.67	3.67	5.23	7.07	3.87	.54
INSPECTION	9.51	8.39	7.99	8.49	8.88	.66
ADMIN SUPER	24.85	27.68	29.88	25.00	27.17	3.05
ADMIN WORKER	12.78	13.94	13.88	9.66	13.31	.72
PARADES/DRILL	1.47	.97	.76	1.31	1.14	.37
SUPPLY	4.00	3.60	.52	.19	3.53	.84

71

100

101

ANALYSIS APPROACH "B"

TO DETERMINE WHAT EFFECT VIOLATION OF TA ASSUMPTIONS HAD UPON THE CLUSTERING PROCESS.

1. OVERLAP GROUP AND DIAGRAM PROGRAMS USED
2. THIRTYNINE STAGES IDENTIFIED FOR STUDY
3. OCCUPATIONAL FIELD ANALYSIS
   A. SPECIAL JOB DESCRIPTIONS BY OCCUPATIONAL FIELD
   B. % OF TIME TASKS REPRESENT

DUTY AREAS IN WHICH MOST MEMBERS WERE PERFORMING

DUTY AREA	POPULATION	PERCENTAGE
1. ADMIN SUPERVISOR	1039	74
2. COUNSELLING	191	14
3. TRAINING	146	10
4. GUARD/CLASSIFIED MATERIAL	28	2

SNCO DIAGRAM

PART 1

659 SNCOs

ADMIN SUPERVISORS DUTY AREA

E-8--E-9

MISC MOSs

TASKS WITH SOFT VERBS

130 TASKS PERFORMED (AVG)

SUPERVISORS

OVERLAP 34 TO 55%

PART II

833 SNCOS

ADMIN SUPERVISORS

TRAINING

COUNSELLING

GUARD/CLASSIFIED MATERIAL

E-6 & E-7

SPECIFIC OF FOR EACH GROUP

TASKS WITH HARD VERBS

40 TASKS PERFORMED (AVG)

WORKER SUPERVISORS

OVERLAP 16-25%

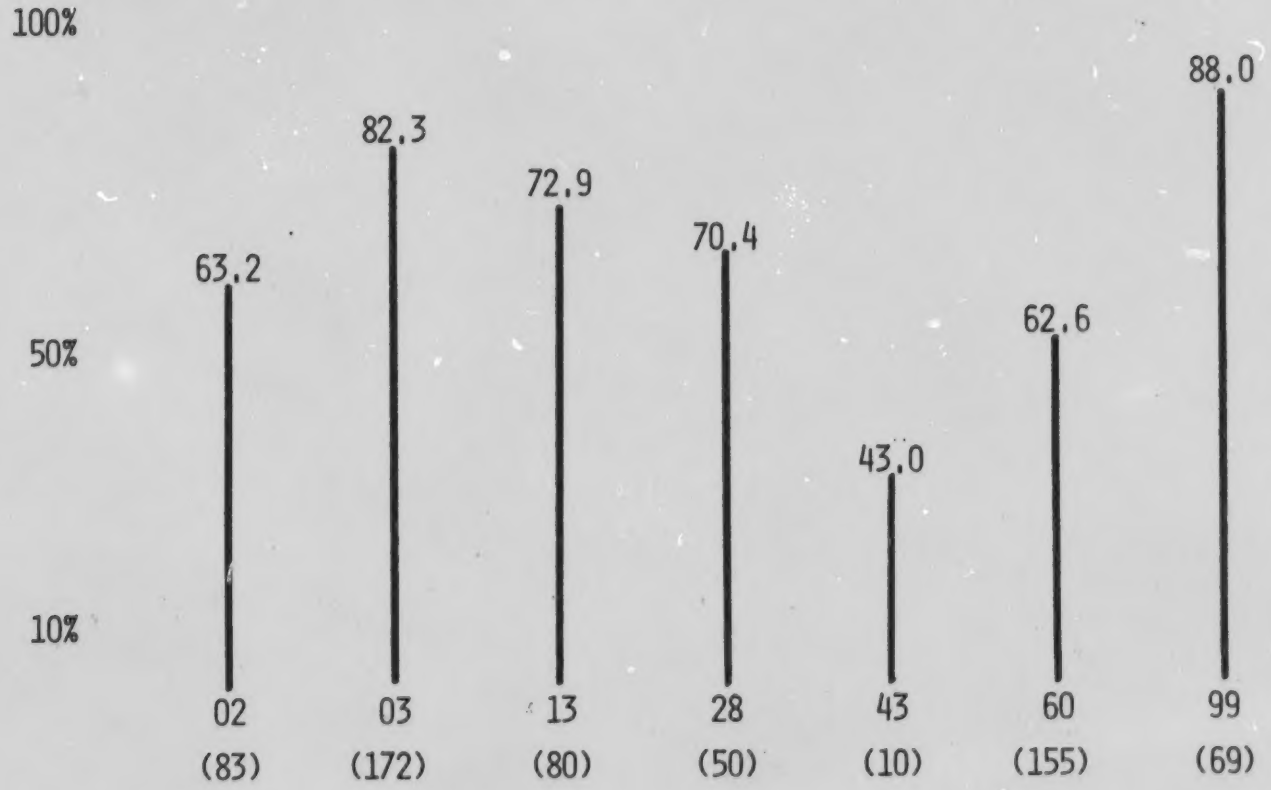
% OF JOB

THE TASKS YOU SELECTED ABOVE REPRESENT WHAT TOTAL PERCENT OF TIME IN YOUR PRESENT JOB?

_____%

% OF JOB (TIME ESTIMATES)

PERCENT TIME	SNCOs
100%	233
70-100%	1036
0-25%	106


% OF JOB BY OCCUPATIONAL FIELD
100%
50%
10%
63.2
82.3
72.9
70.4
43.0
62.6
88.0
02
(83)
03
(172)
13
(80)
28
(50)
43
(10)
60
(155)
99
(69)

77

CONCLUSIONS

1. ALTHOUGH SOME BASIC ASSUMPTIONS ABOUT THE TA METHODOLOGY WERE VIOLATED THE FEASIBILITY OF ADDRESSING COMMON MANAGERIAL/SUPERVISORY TASKS AND NOT ACCOUNTING FOR 100% OF A MARINES JOB STILL PRODUCED WORTHWHILE RESULTS FOR THE SNCOA.

CONCLUSIONS (CONT)

2. THIS STUDY WILL PROVIDE A POINT OF DEPARTURE FOR ATTEMPTING SIMILIAR STUDIES FOR OFFICER SCHOOLS SUCH AS AMPHIBIOUS WARFARE SCHOOL AND COMMAND AND STAFF COLLEGE.

SIMULATION OF COMBAT AS A MEASURE OF INDIVIDUAL PROFICIENCY

A Paper Prepared for the
18th Annual Conference of the
Military Testing Association

October, 1976

by

Robert Pulliam

and

Boyd L. Mathers

I - Introduction

This paper deals with the potential of models, simulations and games (MSGs) in human measurement. In a first part we will comment on the military uses of simulation, and suggest that it has potential as a tool for assessing individual proficiency. To support that suggestion we will review a 1974 experiment, in which the live simulation of small unit engagements was used to measure individual combat skills, with apparently encouraging results.

In a second part we will suggest that a methodology is needed for applying MSG techniques in measurement. The field of military gaming has been marked by advocacy, at the expense of scientific method and empirical validation. We will suggest some basic structual features of MSGs which will affect their application to human factors research, and to measurement in particular.

II - The Potential of Simulation

The only valid criterion against which to judge a test or training program is its application to performance of a real mission. But in peacetime, military organizations do not exercise their primary mission of fighting. Classically, they have prepared for war as best they could, by exercises such as drill and marksmanship. Occassionally there have been efforts to simulate warfare, either as a means to study its characteristics, or to train leaders and men.

In spite of an experience which dates to the dawn of human record, military maneuvers remain poor replicas of reality. And from the earliest times there have been efforts to construct games (such as chess), simulations, and predictive models to study and train for war. These models, simulations and games share the characteristic that they are selective abstractions of reality. We will follow the practice of Shubik and Miller (1971), in using the abbreviation MSG to refer to models, simulations, games and similar creations.

The Uses of MSGs

Chess, of course, was the original MSG. It was assumed to be a representation of war in the seperate forms developed in Europe, China and Japan. More vigorous, if not more convincing as microcosms of war, were the various traditions of tournament and swordplay. Formal war gaming emerged when strategic games were played in Berlin in 1830;

reportedly these games were applied both to analyze strategic options, and to assess the ability of officers to command. Prussian success in the Franco-Prussian War was widely attributed to these games, with the result that by the end of the 19th century use of SMGs was universal among the European powers (McHugh, 1966).

By 1896 tactical games were in use at the Naval War College. They played an important part in both German and Japanese planning for WWII, and following that war the U. S. Joint Chiefs of Staff established what became the Studies Analysis and Gaming Agency. Recent practice has, of course, been typified by elaborate computer modelling.

Modern military gaming has for the most part been in the tradition of systems analysis, and employed for the analysis of political, strategic or tactical options (Shubik and Brewer, 1972). To some extent it has been applied in training. In addition to its military uses, it has been used to predict outcomes of alternative strategies in economics and business. In addition we know that moderately sophisticated mechanical trainers exist to teach navigation, aircrew skills, vehicle operation and electronic troubleshooting. These devices represent an engineering-centered line of effort, superficially different from that of war gaming. But they belong to the family of MSGs, and from a behavioral point of view they are similar in their nature and effects. What we notice is that there is no discoverable literature on the use of simulation in testing.

We suggest that this indicates an unexploited opportunity. We will argue that simulation is unexplored partly because the few efforts in that direction have not led to the development of any general theory, mathematics or vocabulary for applying MSGs in testing and related behavioral research.

The practice of military gaming and of systems analysis has applied elaborate computer simulations, and has a certain basis in mathematics. The classical treatment of games by Von Neumann and Morgenstern (1944) is suplemented by a substantial literature on the modelling of systems in logistics, business practice, industrial production (Mize and Cox, 1964), and biology of organisms (AIBS, 1975). These models make it possible to experiment using models of systems, where it is not desirable to experiment upon the systems themselves. But in the case of military and international gaming, there is a suprising poverty of basic theory to support the operational work, as is noted by both McHugh (1966) and Shubik and Brewer (1972). A science of SMGs is possible. Such a science is anticipated by the thorough analyses which underly flight simulators at the engineering level, and comparable behavioral theory which was developed to support flight simulators for the NASA manned space program. Now similar basic science and methodology is needed concerning the process of simulation itself. Such a science could make it possible to:

o Predict the consequences of general SMG strategies, such as choosing a board game vs. a computer simulation.

o Recognize what tactical options are open, in designing a particular simulation or game.

o Recognize the behavioral consequences of any design tactic.

o Recognize the measurement options which are useful to record and analyze behavior in simulations.

It has yet to be demonstrated that simulation can yield useful data concerning individual performance. To that end, we will review an experiment previously reported in 1974.

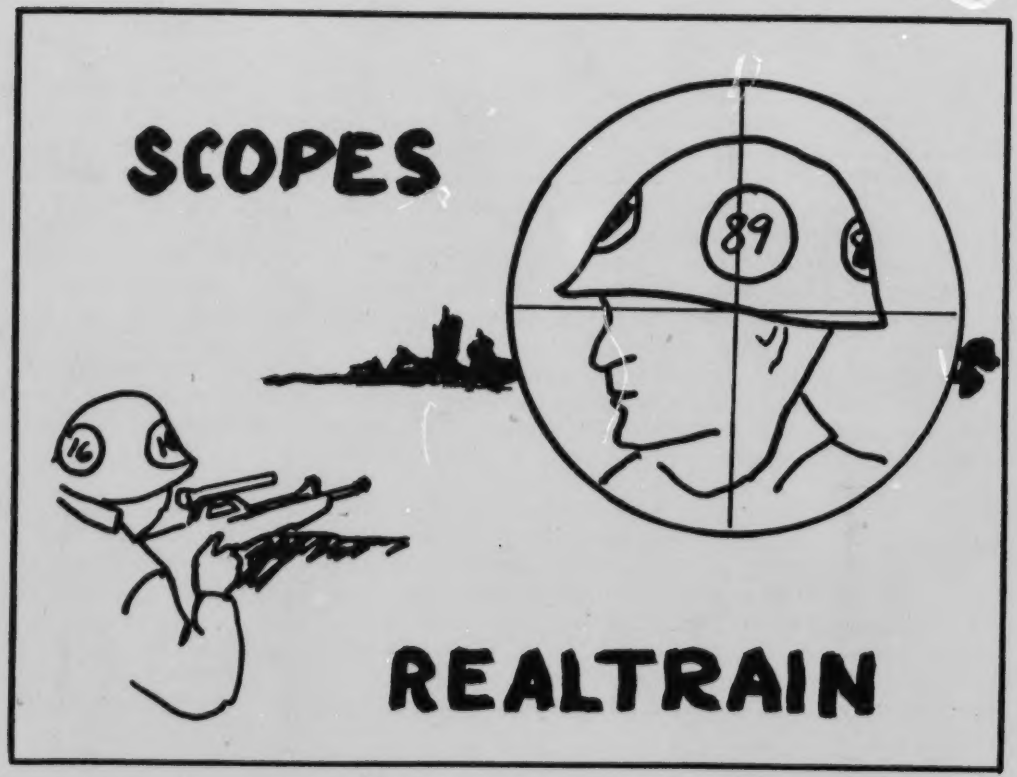
The SCOPES/REALTRAIN Experiment

A potential of combat games to assess the proficiency of individual soldiers is at least suggested by an experiment which we will now review.

The Army Research Institute has developed a family of techniques for simulating combat at the small-unit level. These games pit one squad or platoon against another in moderately realistic opposition exercises, fighting in real time on real terrain. As developed the games require no elaborate simulators or scoring equipment, but use mostly the normal resources of a battalion. They are free-play, two party, negative-sum games with no prearranged scenario; the game is fully defined by its scoring technique. The outcomes, win or loose, depend realistically on the battle decisions of small unit commanders and the actions of their men.

These games have been implemented as training under the names SCOPES and REALTRAIN. They consist essentially of systems for simulating weapons effects and scoring casualties in the field. They provide simple means for identifying and eliminating players who are hit by simulated small arms fire, indirect fire, or the effects of other weapons including mines and booby traps, hand-held antiarmor weapons, and tank guns. They can score interchanges of fire between armor and footsoldiers. We will not have time to describe them in detail, but for those who have not heard of them before, it may help to explain the system for scoring rifle fire.

As with other weapons, the method requires only the simplest technology. Each participating infantryman carries a two digit number on his helmet cover, not readable with the naked eye beyond a few yards. The number is not known to the "enemy". Each rifleman has an inexpensive scope mounted on his M-16 rifle. As the opponents maneuver against each


SCOPES
89
REALTRAIN

(4) A peer ranking procedure required each participant to list his fellow soldiers in rank order, according to the " . . . order you would like to have by your side . . ." in combat.

(5) Finally there was a simulation sequence. There were 6 situational combat simulations: an assault, a defense, two kinds of patrol activity, an anti-tank exercise and a mechanized movement exercise. Individual performance was scored on the basis of hits made, on the initiative of the member concerned, including hits on armor and as a result of individual initiative in requesting artillery fire. We will call this measure the Combat Simulation.

(6) Finally, a composite score was computed by summing the raw scores for the 5 measures just described, with equal weighting.

Findings. These tests were administered to 9 NCOs in leadership positions (MOS 11B40) and to 26 riflemen (MOS 11B20).

Table 1 shows all cross correlations for the 5 basic measures, and correlations of each measure with the composite score. From left to right you will note:

- o Oral Knowledge
- o Skill Performance
- o Controller Ratings
- o Peer Rankings
- o Combat Simulation
- o The Composite Score

Each datum in the matrix is a correlation (Pearson r) between the two measures which intersect at that point. Each intercorrelation appears twice in the matrix. These data are for NCOs. There are at least three things to note:

First, notice the surprisingly low relationship between all other tests and performance tests for combat skills, as measured individually and not in a combat context. These presumably practical tests correlated negatively with other measures, and at near-chance levels of significance. To the extent that these findings are replicable, they suggest that measurement of combat skills using an SQT strategy could be hazardous.

Peer ratings, on the other hand, were significant and positively correlated with the other remaining measures. So were simulations.

Table 1

NCO TEST PERFORMANCE (1)
Intercorrelations

MEASURE	Oral Knowledge	Skill Performance	Controller Ratings	Peer Rankings	Simulation	Composite Score
Oral Knowledge	X X	.186	.460*	.689*	.588*	.808**
Skill Performance	.186	X X	-.285	-.389	-.264	-.017
Controller Ratings	.460	-.285	X X	.844**	.782*	.727*
Peer Rankings	.689*	-.389	.844**	X X	.669*	.697*
Simulation	.588*	-.264	.782*	.669*	X X	.942**
Composite	.808**	-.017	.727*	.697*	.942**	X X

* Significant at .05 level.
** Significant at .01 level.
(1) After Hayes, Griffin and Mathers (1974).

These data may be seen more clearly in the simplified format of Table 2. This chart contains the same data, clarified by eliminating the redundant cells, and by omitting all correlations of less than .20. All the remaining data are significant at he .05 level, except those marked with an asterix, which reach the .01 level.

It will be apparent that, although peer rankings were a valuable measure, combat simulation was the most dependably intercorrelated with other measures (except skill performance), and was the most reliable predictor of the composite score. Findings for soldiers were similar to those for NCOs.

Table 3 shows the same data for 26 infantrymen.

In this case, only peer rankings and combat simulation produced interesting correlations. We observe that:

(1) Controller ratings were of no value. This is presumably due to the fact that exercise controllers interact primarily with NCOs rather than soldiers. They have difficulty recognizing individual soldiers during an exercise, and even greater difficulty remembering what they do.

(2) All the other measures are correlated with the composite score (bottom line on the table) at a level only slightly above that caused by the contribution of their own value to the composite total.

(3) Skill performance tests, as expected, appeared more useful in this assessment of soldiers than in assessing NCOs in leadership positions. They were significantly correlated with peer rankings, but peer rankings were better predictors of the composite score.

(4) Combat simulation was significantly correlated with oral knowledge and with peer rankings, the two other tests which, at this point in analysis, were considered to have the highest presumption of validity. Simulation was slightly correlated (.174) with skill performance, a measure of questionable validity, and not correlated (-.015) with controller ratings, which were also considered questionable.

(4) Combat simulation was able to predict the composite score at the very high correlation of .927 (Pearson r).

Summary

We can make few final conclusions from this experiment, which was exploratory and based on a small (35) population. The Army chose not to follow up on it for practical reasons. For one thing, combat simulations are inherently non-standard. Each engagement proceeds differently, and tests are supposed to depend on replicated conditions of performance.

Table 2

NCO TEST PERFORMANCE [1]

Intercorrelations [2]

MEASURE	Oral Knowledge	Skill Performance	Controller Ratings	Peer Rankings	Combat Simulation
Skill Performance	- -				
Controller Ratings	.460	- -			
Peer Rankings	.689	- -	.844*		
Combat Simulation	.588	- -	.782	.699	
Composite	.808*	- -	.727	.697	.942*

(1) After Hayes, Griffin and Mathers (1974).
(2) Only correlations greater than .02 are shown.
* Significant at the .01 level. All other data shown are significant at .05.

Table 3

SOLDIER TEST PERFORMANCE [1]

Intercorrelations [2]

MEASURE	Oral Knowledge	Skill Performance	Controller Ratings	Peer Rankings	Combat Simulation
Skill Performance	- -				
Controller Ratings	- -	- -			
Peer Rankings	- -	.456	- -		
Combat Simulation	.364	- -	- -	.361	
Composite	.492	.412	- -	.420	.927*

(1) After Hayes, Griffin and Mathers (1974).
(2) Only correlations greater than .02 are shown.
* Significant at the .01 level. All other data shown are significant at .05.

Furthermore combat games are credible as measures of unit performance, but less credible for measurement of individuals. No way has been shown to compensate the man who is ordered into a place of high tactical hazard, or one who takes risks in the interest of the mission. Finally, even the SCOPES exercises are imperfectly equal to combat, and costly in time and effort. So the Army chose to develop the simulation techniques used at Ft. Benning not as performance tests, but for training.

Neverless the reported data suggest that simulation did very well as a test of individual competence. The data suggest two recommendations:

(1) That tests of performance for single skills should be interpreted only with great caution as suggesting combat competence.

(2) That individual proficiency might be measured in the context of a realistic battle exercise, and that such data might be superior to any other measures yet proposed.

III Towards a Theory of Simulation

We have reviewed a 1964 experiment as evidence that simulation is worth exploring. SMGs have a potential, not yet proven, to measure individual differences in complex activities, where conventional measurement is not successful. We wish further to suggest that simulation as a technique for measurement has been little explored in part because there is little precedent for doing so. There is no appreciable literature on the subject. There is no body of proven method, no vocabulary, no applied mathematics, and no developed theory of simulation applied to human factors research.

Conventional Testing

In contrast, normative testing benefits from a rich precedent. Over the past 100 years we have developed a methodology, a statistical science, and a profession to advance the measurement of human factors examined, deliberately, out of operational context and as seperate dimentions.

In conventional testing, scores are a function of several responses to individual test items. We are able to sharpen the resolving power of items by statistical treatment, and to quantify reliability, validity and response clusters. We are now very good at that.

In fact we are so good that it has become a habit. The tradition is so strong that, when faced with a need for information, we tend automatically to apply the cookbook for item-based normative tests. Such tests are elegant measures of knowledge, and of those other human variables which are fitted to paper-and-pencil tests. Actually we are not very good at measuring most real-life task performance, nor aptitudes other than those

related to general intelligence and academic learning. We are just discovering the weakness of our methods for the measurement of job content (Christal, 1974), and job performance (Hirshfeld, et. al., 1976).

We will offer some suggestions leading to a general understanding of what simulation is, and what it can do for measurement. These suggestions place SMGs in context with related sciences, and imply a general theory of simulation. They are not considered complete or final, but are offered to stimulate other thought directed toward a disciplined applied science for the use of SMGs in human factors research.

We begin by observing that combat simulations and war games belong to a general class of phenomena, related approximately as shown by Table 4: Refering to that table, A model is a selective abstraction of reality - a subsystem from the real world, in one of the following forms:

o A replica is a thing reproduced as exactly as practicable, like a copy of a 1912 Mercedes.

o A physical abstraction is a creation which reflects only selected parts of a real system. A ship model reproduces the function and shape of the real thing, but not the size. A molecular model reproduces the form, but not the size or function of its object. A combat game simulates the sights and sounds of warfare, but not the injuries. Formal scientific experiments are based on physical abstractions, or on algorithms.

o An algorithm is an equation representing selected interactions in nature. It expresses a physical abstraction in terms of dimention, quantity or probability.

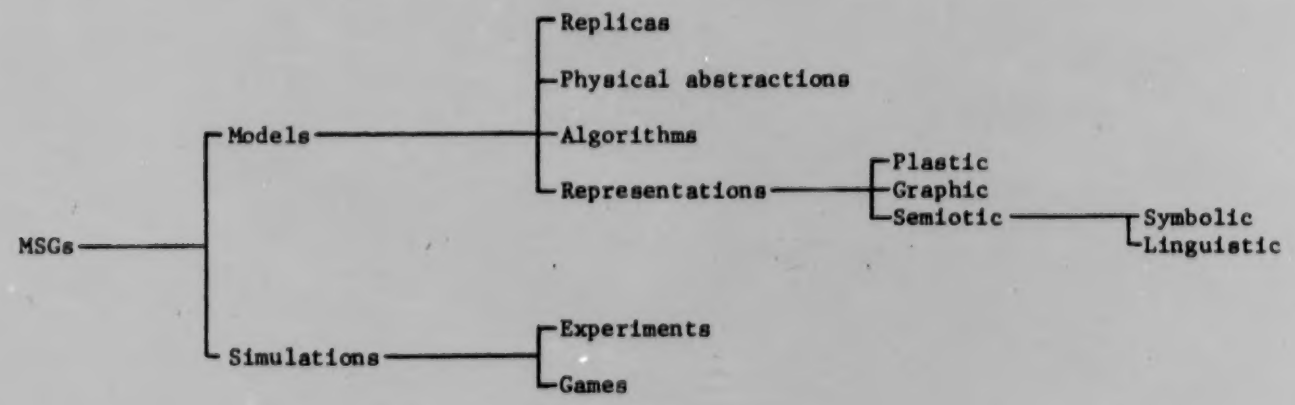
o A representation copies a subsystem of reality, not in its literal form or material but in terms of representative shapes or sculptures (plastic), pictures or maps (graphic), suggestive or conventional symbols (symbolic) or language (linguistic).

Models are static creations: simulation is a process. Simulation is the use of models to represent selected events in nature dynamically. Simulations are created by setting models in motion (as in a physics experiment), by performing an activity under control of a model (as in a SCOPES exercise), or by applying a model to observation of reality (as in identifying an insect to genus).

Models are constructs, either representational or theoretical. Simulations are specific exercises of a model, usually applied for information related purposes such as to study behavior of a system or to train operators. Occasionally models are applied for non-information purposes: If you have a replica of a 1912 Mercedes, you might ride it to the golf course.

Table 4

Semantic Tree for MSG Terms



Simulations (still referring to Table 4) include the major categories of experiments and games.

o Experiments include all simulations of reality intended to discover knowledge and to predict the future behavior of natural and man-made systems. Wargaming is clearly in this category when used to evaluate weapons and tactics, or to explore strategic options.

o Games include simulations in which one or more human decision makers interact with the model while it is running. Games include an adversary relationship between the players, the system, and chance, with information processing by the players as the event of central interest. Games may provide full or incomplete information, and offer a gain or loss (payoff) to the players. Psychologically, game play is imperfectly understood.

Combat games such as SCOPES and REALTRAIN combine elements of both experiments and games. Each time a player makes a decision to move, fire or signal, he initiates an experiment. The model predicts what the outcome of such an action might be in battle. Players and leaders proceed by a series of tactical experiments, performed upon the situation and the enemy. Hopefully they learn thereby. But combat simulations are games in the sense that they involve the interaction of adversary decisions, with incomplete information, a payoff, and the psychological consequences of a contest. Whether driving an antique car to the golf course is a simulation, or a game, we leave the reader to ponder.

A Definition

We are now equipped to suggest a definition:

> Simulation is the process by which man intervenes in events within nature, using a model, to produce a microcosm of events which are selectively representative of events historically observed, and presumably representative of possible future events in nature.

Simulation can also be defined by its uses: It is the use of models dynamically to:

- o Discover information about nature.
- o Predict future behavior of a system.
- o Train human operators of systems.
- o Measure human responses in systems.

Simulation is not a gimmick. It belongs to the mainstream of science as the second of two methods for disciplined inquiry defined by Aristotle.

The first tool of science is logic; the second is systematized observation. Simulation is that subcategory of observation in which human intervention controls the conditions under which nature is observed. It is identical with experimental science.

SMGs include classical experiments, but also permit cases in which models are permitted to run with large numbers of free variables. Usual experiments in the natural and behavioral sciences presume that all variables which can affect the outcome, except two, are held constant. The control variable is then manipulated while the effect on the dependent variable is observed. This process fits our definition of simulation, as a controlled representation of reality. However the simulations which are of most interest to us here, as in combat simulation, have many free-running variables. More than one variable may be manipulated, and more than one measured. Alternatively we may may observe only an outcome (win or loose).

The Relationship of Model and System

Any model may be characterized by its relation to reality, in the manner suggested by the Venn diagram of Figure 1:

Where a system is represented by the field S, a field which contains all characteristics or variables of the system, let the field (M) represent the model, and include all characteristics or variables of the model.

Then the model is fully defined in its relation to the system when we know:

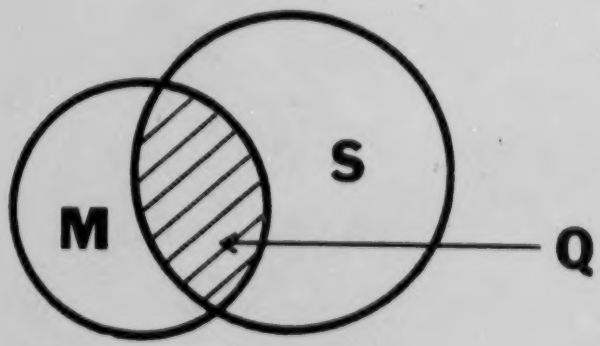
- What characteristics of (S) have been captured in the field of representation (Q), at the intersection of (S) and (M).
- What characteristics of (S) have been omitted, either unavoidably or by choice, and are in the field (S) -(Q).
- What characteristics the model (M) incorporates which are not characteristics of the system. This field (M) -(Q) may include undesired artifacts or imperfections, or desired interventions such as an ability to stop the clock, to change scales of time or space, to manipulate variables, or a capability to measure and score.

With this understanding we can quantify the adequacy and reliability of the model:

- Adequacy is the degree to which the model represents the system fully, and is is a function of the area of Q divided by the area of (S). Adequacy therefore approaches 1.0 as the model becomes more

1664

Figure 1
Relation of Model to System



Q = Representation

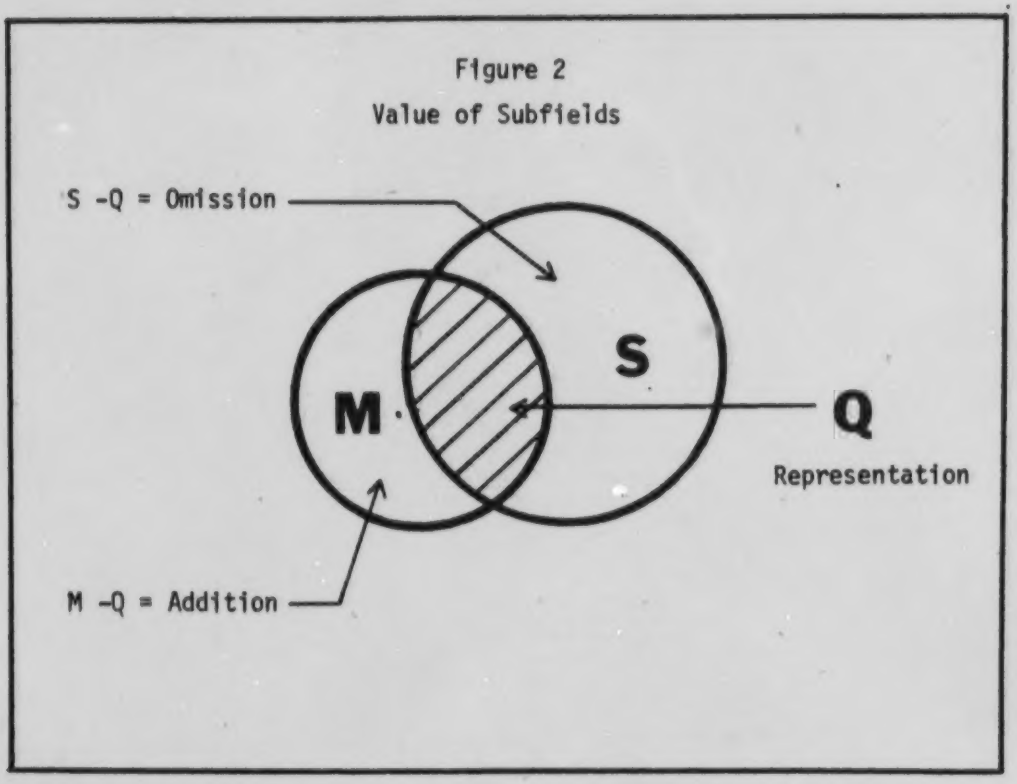


Figure 2
Value of Subfields
S -Q = Omission
S
M
Q
Representation
M -Q = Addition

representative, until at 1.0 the model is in fact a replica. (Figure 2).

o Reliability is the degree to which variables not derived from the system intrude upon the behavior of the model, and is a function of the area of (Q) divided by that of (M). In the ideal case (M) would be fully contained within (S), and equal in area to (Q). (Figure 3).

A science for the use of MSGs in human factors research must presumably depend on selective control of the content of field (Q), and of the fields (S) -(Q) and (M) -(Q), considering in each case the behavioral consequences rather than the literalness of the model.

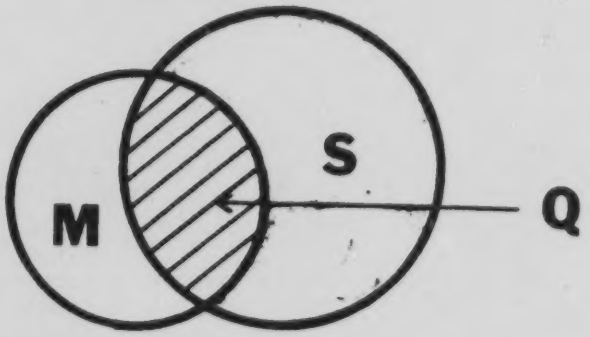
Management of the Model

Within each segment of the Venn figures, we need a theory and vocabulary to assist us in deciding whether to include or exclude a variable from the field, based on the purpose we intend. In general, we suggest that existing training simulators will be found to be over inclusive. Simulators usually have been designed either by teachers or by engineers. They tend to be near replicas of reality.

For example, we have seen circuit analysis trainers for troubleshooting electronics which use real hardware, with added instrumentation. In specific cases they are used to illustrate only one or two stimulus-response relationships which very primitive equipment could represent as well. In one USAF school, students and staff were not allowed to touch the available real equipment, because it was too valuable. We have observed tank-turret trainers in a contractor facility which were a cut-out section of a real tank. It appeared uneconomical to use this 7 ton simulator for the purpose of teaching men to troubleshoot an electrical system weighing not over 40 pounds. It would be equally unnecessary to use a high level of replication to measure maintenance skill.

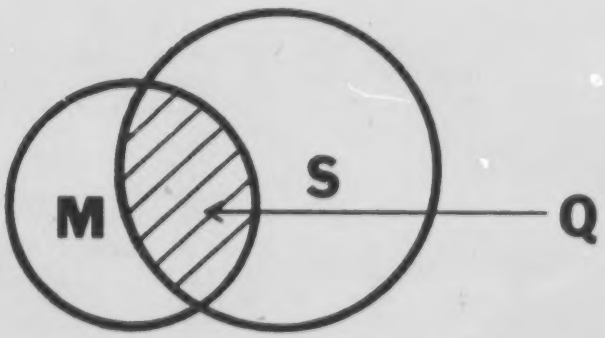
We noted earlier that the field (S) -(Q) contains characteristics not included in the model. The practical difficulty in policing the content of this field is that we clearly do not know all the characteristics of any system. In combat simulation there are many suspected variables which we cannot isolate or define. For example we cannot model the interactions of weapons skill, physical stamina, perceptual alertness, imagination, intelligence, interpersonal signalling, tactics, situation and terrain. It must not be inferred that simulation fails when these effects are not understood. In fact we have the advantage that we can fall back on constructing a simulation with high physical fidelity, then test it for psychological fidelity, and proceed with a high degree of confidence that the outcomes of these many interactions will be near to the outcomes of reality. The advantage of simulation is that it is not necessary to be able to define, track and measure an interaction in

Figure 3
Adequacy



Adequacy $a = \frac{\mu (Q)}{\mu (S)}$

Figure 4
Reliability



Reliability $r = \frac{\mu\ (Q)}{\mu\ (M)}$

order to include it.

As a matter of fact it is an advantage of simulation that, if tests of fidelity are met, there is a high presumption that many unknown variables are in fact withing the field of representation (Q). Contrast this with the case of normative, item based tests. Here there is a high presumtion that, along with undesired and interfering variables, the process of item analysis will screen out all variables or attributes which may contribute to performance, but which have not been defined.

Rules for Application

Finally we will suggest some principles to keep in mind when using simulations in measurement or in training. They will be listed without any systematic order:

(1) Simulations are effective when they are cost effective in terms of their time and space scale. One advantage which the simulator can offer is an ability to cover more experience in less time, or to stop the clock if it is useful. The same principle applies to space. A half-scale simulator may measure mechanical behaviors more effectively than real equipment.

(2) The key issue in designing the model is to control the content of field (Q). In doing so, maximizing the area of (Q) is not desirable. Simulation will be least expensive and most effective when only the important attributes of the system are treated by the model, to the exclusion of others.

(3) In specifying the required content of field (Q), the issue is not what variables are treated from an observer's point of view, but whether the behaviorally significant features of reality are included. Thus we should ask about any possible variable: What are the cues or stimuli to which the subject reacts? What are the conditions which affect his perception and reactions? What behaviorally important response repertoire is possible? What are the normal and behaviorally necessary feedbacks and reinforcements which accompany each response?

(4) It is particularly important to observe the information display to which subjects respond. That display should contain the appropriate cues for action, within a field of appropriate distractors, and require a realistic perceptual and intellectual search prior to action. Human behavior (especially in combat) is not based on gross reactions to Skinnerian stimuli, but on a heuristic search for information. Cues are actively sought in a field of noise, and are rarely unequivocal.

(5) The outcome of a game or exercise must be realistically the conseqence of decisions made by players. This sounds obvious, but is

a rule frequently disregarded, either in the interest of simple game mechanics or (at the other extreme) in the interest of unneeded fidelity.

(6) Simulation should do no harm. The field (M) -(Q) should be searched for undesirable learning or events which might interfere with the measurements intended. In particular, game designers must ensure that the subject is not required to learn enabling skills - skills required solely by the game - which are either an excessive learning load or potentially likely to interfere with desired behavior.

In Summary

There is reason to believe that simulation techniques have potential to assist in human factors measurement, especially where there are many interacting variables, where existing measures do not work, or where those measures are of questionable validity. Simulation has been demonstrated as a means of individual measurement in small-unit combat, and would appear to have potential as a practical measure of battle competence. Its use might be explored in other interactive situations, such as:

- o Tests of organizational behavior (as with in-box exercises).
- o Individual performance under stress, or high signal to noise conditions (AC&W, Antisubmarine operations).
- o Interviewer effectiveness (interrogation, recruiting).
- o Foreign language proficiency under real-life conditions (the cocktail party task).
- o Leader selection.
- o Communications network skills.

In this paper we have taken an advocacy position. But we do not suggest that simulation is a magic solution for complex testing problems. We do believe that it is a technique with potential, and one which has been largely unused. In particular we affirm that it is the only known tool for measuring human performance under certain real conditions. The situations in which it applies include cases where the number of interacting behavioral variables is high, and where important task variables resist definition.

REFERENCES

AIBS (1975) Proceedings of the American Institute of Biological Sciences, Session on Models and Micrcosms. Oregon State University, 1975.

Aristotle, Organon.

Christal (1974) Raymond F. Christal, The Air Force occupational research program. San Antonio, Texas: AFHRL (1974).

Hayes, et. al. (1974) John Hayes, Glenn R. Griffin and Boyd L. Mathers, Development of performance-based tests for combat arms skills. Final report contract DAHC19-73-C-0016. Arlington, Va: ARI (1974).

Hirshfeld, et. al. (1976) Stephen F. Hirshfeld, Douglas L. Young and Milton H. Maier, Procedures for Validating Skill Qualification Tests. Arlington, Va.: ARI (1976).

McHugh (1966) F. J. McHugh, Fundamentals of War Gaming. Newport, N. Y.: Naval War College (1966).

Mize and Cox (1968) Joe H. Mize and J. Grady Cox, Essentials of Simulation. Englewood Cliffs, N. J.: Prentice Hall (1968).

Shriver and Pulliam (1976) Edgar L. Shriver and Robert Pulliam, "Ratio-of-forces techniques in the scoring of combat simulation." Proceedings of the Military Testing Association. Pensacola, Fla.: (1976).

Shubik and Brewer (1972) Martin Shubik and Garry D. Brewer, Models, Simulations and Games - A Survey. Final report contract DAHC15-67-C -0141. Santa Monica, Cal.: ARPA (1972).

Von Neumann and Morgenstern (1944) John Von Neumann and Oskar Morgenstern, Theory of Games and Economic Behavior. Princeton, N. J.: Princeton University Press (1944).

MULTIPLE CHOICE TESTS: HOW MANY ALTERNATIVES?

by

Major Ronald G. Swanson
Academic Instructor School
Maxwell Air Force Base, Alabama

Tversky (1964) considered tests to be sequences of n choice points (items) with a alternatives per choice point, where the total number of alternatives is a constant c (c = na). Fixing c requires that test length increase to compensate for the reduced number of alternatives per question and vice versa.

Grier (1975) examined how the number of alternatives (a) might affect expected test reliability. The highest reliabilities were expected when three alternatives per item were used. Grier concluded that for any given c value, three alternative test items would maximize: (a) expected test reliability, (b) the power of a test, (c) the discrimination capacity of the test, and (d) the uncertainty index. Other supposed advantages of three-alternative items included relative ease in item construction, a likely gain in test efficiency, more comprehensiveness, and little, if any, increase in the amount of time required to administer the test.

The present study was conducted at the Air Force Academic Instructor School (AIS) to determine if Grier's theoretical position on three-alternative test items would be of practical use in an applied setting. The subjects were 489 U.S. Air Force students in AIS Classes 76-A, B, C and D. The test used was the AIS Academic Post-Test which was given in two forms—the normal 75 item four-alternative test and an equivalent 100 item three-alternative test (note constant c of 300).

The results indicated that the three-alternative test had consistently higher reliabilities, higher average differentiation indexes, and took little additional time to administer (indicating higher test efficiency). Three alternative items were considered easier to construct and resulted in tests that were more comprehensive. It was concluded that three-alternative multiple-choice tests should be closely examined by military testors in light of the results obtained.

MULTIPLE-CHOICE TESTS: HOW MANY ALTERNATIVES:

by

Major Ronald G. Swanson
Academic Instructor School
Maxwell Air Force Base, Alabama

This study supports the hypothesis put forth in a previous study that three-alternative multiple-choice test items are more efficient than items with four or five-alternatives.

The most common test item type used in military (or for that matter, civilian) testing is the multiple-choice test item. Reasons for this are numerous and include objective scoring, ability to test all of the cognitive levels of learning (Bloom, 1956), student preference for selection type test items, and relative efficiency over other item types.

While multiple-choice items are widely used in military testing, there appears to be some misunderstanding about them. Many schools and commands direct item writers to provide four, and only four, alternatives per multiple-choice item. Sometimes five alternatives are demanded under the guise of obtaining higher test reliabilities. This demand is based on the assumption that all of the item alternatives will be good ones, not always (or even usually) the case.

Where did this notion come from that four or five-alternatives are optimal in multiple-choice items? Most test constructors claim that more alternatives per item will result in higher test reliabilities due to increased test length. While this assertion is theoretically true, is it valid in applied settings where real tests must be written by real people to be taken by real students? Does the use of four or more alternatives per item have beneficial or detrimental effects in regard to other characteristics of evaluation such as differentiation and comprehensiveness?

Tversky (1964) considered tests to be sequences of n choice points (items) with a alternatives per choice point, where the total number of alternatives is a constant c ($c = na$). Fixing c required that test length increase to compensate for the reduced number of alternatives per question and vice versa. The use of three alternatives per item was found to maximize three criteria: the power of the test, defined as one minus the probability of getting a perfect score by chance; the discrimination capacity of the test, defined as the number of possible response patterns the test can distinguish between; and the uncertainty of index, a measure of the information gained from using the test. Grier (1975) provided a tabular representation of a simple example of the set of tests that can be formed from c=12 alternatives and the discrimination capacity of each (see Table 1).

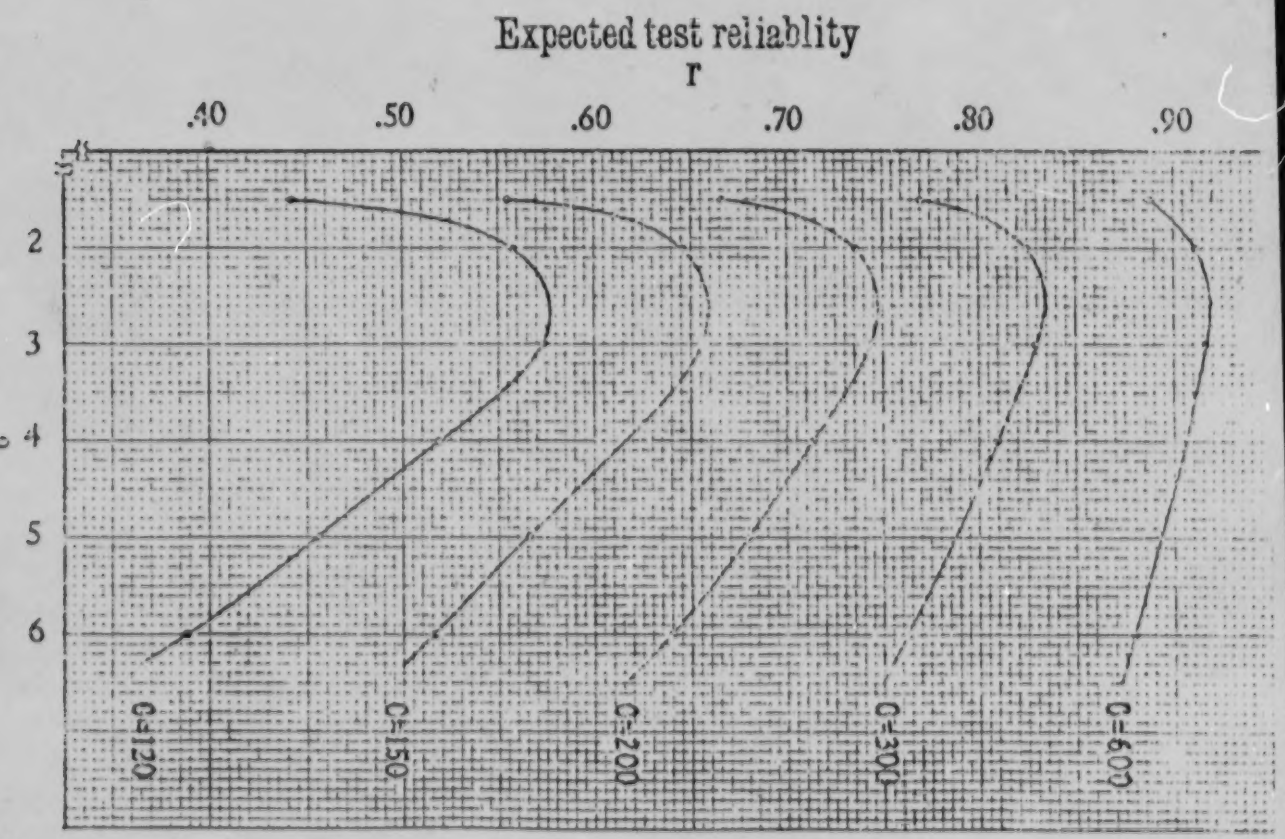
Table 1
Test Sets from c=12 and Resultant Discrimination Capacity

Number of Questions (n)	Number of Alternatives each (a)	c=na Constant	Discrimination Capacity
6	2	12	64
4	3	12	81
3	4	12	64
2	6	12	36
1	12	12	12

(from Grier, 1975)

Grier also examined how the number of alternatives (a) might affect expected test reliability. The highest reliabilities were expected when close to three alternatives per item were used (See Figure 1).

Figure 1
Expected Reliability (r) of a Test of Size c as a Function of the number of alternatives (a).



(from Grier, 1975)

Grier concluded that for any given c value, three alternative test items would maximize: (a) expected test reliability, (b) the power of a test, (c) the discrimination capacity of a test, and (d) the uncertainty index. Other supposed advantages of three-alternative items included relative ease in item construction, a likely gain in test efficiency, more comprehensiveness, and little, if any, increase in the amount of time required to administer the test.

One important facet of Grier's suggestions which was incorporated into the present study was the maintenance of a constant c value. Ramos and Stern (1973) examined four and five-alternative items and found the five-alternative items superior in reliability, but this could have been expected since they did not increase the number of four-alternative items in order to maintain a constant c.

The present study was conducted at the Air Force Academic Instructor School (AIS) to determine if Grier's theoretical position on three-alternative test items would be of practical use in an applied setting.

Method

Subjects

The subjects were 489 AIS students ranging in age from 20 to 56, and in rank from sergeant (E-4) to colonel (O-6). These subjects were students attending AIS during classes 76-A, B, C, and D (January to July, 1976), each of which were five weeks in length. Ninety-four percent were male. Classes 76-B and D were predominately officers with graduate degrees; classes 76-A and C were predominately NCOs, some with bachelor degrees.

Instrument

The test used was the AIS Academic Pre/Post Test which normally consists of 75 four-alternative multiple-choice items. An experimental test consisting of 100 three-alternative items was developed (note that c remains constant at 300) which employed the 75 items from the normal test, with one less alternative each, and 25 additional items of approximately equal difficulty.

Procedure

When the students for each class arrived at AIS, they were randomly divided into seminars of eight or nine each. Seminars were then randomly assigned to one of two groups: (a) the experimental group which took the 100 item three-alternative test or (b) the control group which took the 75 item four-alternative test.

Since the AIS Academic test is used in a pre/post design, students take the same test on the first and last full day of class with a five week interval between administrations. The purpose of the pre-test is to establish a base-line performance against which to compare later performance. Hence, on the pre-test, students were told not to guess at questions about which they felt totally ignorant. On the post-test,

guessing was neither encouraged nor discouraged. To avoid comparison problems, it was decided to use only the post-test administration for the purposes of this study. This decision eliminated the guess factor distortion. It also insured that all students were exposed to the same instruction and experiences prior to the academic post-test and theoretically could answer all post-test questions.

The test administrations were timed, including the time required to transfer the answers selected to a computor scan sheet for later analysis. It should be made clear that the academic post-test is not a timed test. More than enough time was available for even the slowest student to finish all items. Hence, there was no more time pressure on students in this study than would normally have been the case.

Results

The test reliabilities (KR21) for each test in each class were calculated and are summarized in Table 2. This analysis reflects consistently higher reliabilities for the 100 item three-alternative test.

Table 2
Summary of Test Reliabilities

Test	76-A	76-B	76-C	76-D
75	.82	.83	.81	.58
100	.86	.88	.85	.76

Mean scores, standard deviations, ease indices, and differentiation indices were calculated for each class and are presented in Table 3.

Table 3
Test Data Summary Table

Class	Test Version	Mean Score	Standard Deviation	Average Ease Index	Average Differentiation Index
76-A	75	46.98	10.70	62.64	.28
(N-109)	100	64.51	11.16	64.49	.36
76-B	75	54.60	9.17	72.80	.31
(N-142)	100	70.49	12.94	70.49	.34
75-C	75	48.66	9.26	64.91	.32
(N-123)	100	65.20	12.12	65.10	.34
76-D	75	56.98	5.66	75.96	.20
(N-115)	100	76.60	8.49	76.58	.24

It should be noted that the average ease indices for both test versions were very similar, indicating that the tests were of equal difficulty. Though the differences in differentiation were small, there was a consistent advantage for the 100 item test.

The minimum, maximum, and mean times were obtained for each test administration and are presented in Table 4.

Table 4
Summary of Time Data by Class in Minutes

Class	Test Version	Minimum Time	Maximum Time	Mean Time	Difference in Mean Time
76-A	75	36.0	85.0	58.1	1.3
	100	30.0	88.0	60.4	
76-B	75	30.0	95.0	57.2	1.8
	100	35.0	88.0	59.0	
76-C	75	30.0	88.0	58.0	6.5
	100	35.0	82.0	64.5	
76-D	75	30.0	88.0	55.5	6.5
	100	35.0	82.0	62.0	
Total Sample	75	30.0	95.0	57.2	4.2
	100	30.0	88.0	61.4	

Table 4 reflects very small mean time differences between test versions. It is also evident that the minimum times were very similar as were the maximum times. However, examination of the maximum time values indicates that the 75 item test acually took longer to take (in most classes) than the 100 item test.

Discussion

The results indicate some distinct advantages for the three-alternative test. Consistently higher reliabilities and differentiation indices are strengths desired in any test, but as Grier has suggested, there are other advantages to three alternative test items.

First, it is usually easier for item writers to come up with two plausible distractors than with three or four, a fact attested to by comments from the item writers involved. The fourth and fifth alternatives are the toughest to come up with and are often the weakest alternatives in an item. In revising test items, it is easy to throw out the poorest distractors in an item and end up with a correct answer and two good alternatives. The inability of test constructors to come up with plausible alternatives also results in the overuse and misuse of such phrases as "all of the above," "none of the above," or obviously wrong answers. However, it should not be concluded on the basis of this study that all four and five alternative items should be discarded. Some questions lend themselves readily to four good alternatives, such as north, south, east and west, or any other logically complete set of four or five answers. Nevertheless, even in these items one or more of the alternatives can usually be eliminated and not seriously affect the validity of the item.

Second, there is an apparent gain in testing efficiency since the additional 25 items in the 100 item test increased the comprehensiveness of the test and took very little additional student time. The time factor may have been the result of students being able to carry the possible responses in their heads and not needing to reread the stem and the alternatives. This increased efficiency is especially apparent in the original 75 test items where the only change was the elimination of some of the alternatives, with stems remaining exactly the same. While

current military austerity programs would seem to make increased testing efficiency very desirable, efficiency in testing is an area of education that has been almost totally ignored.

Another issue that must be dealt with is that of mixing items with different numbers of alternatives in one test. While there is nothing wrong with such a procedure, three important factors must be considered: 1. Grier's hypotheses and the results of the present study indicate that three-alternative items maximize a number of important testing characteristics. The addition of alternatives should be carefully considered in this light. 2. As in any multiple-choice item, the alternatives used must be good distractors. One strength of the three-alternative item is that the distractors are more likely to be good. 3. When mixing items of different length in a new test, the compiler must take into account the *total* number of alternatives to maintain an approximation of the original c value. If this factor is not considered, the students may require more time to complete the new test, thus lowering the test's efficiency.

The results of this study fully supported Grier's hypotheses and should open new vistas in military testing if cross-validated in other testing situations.

REFERENCES

Bloom, B.S. (Ed.) Taxonomy of Educational Objectives. Handbook I: Cognitive Domain. New York: McKay, 1956.

Grier, J.G. The number of alternatives for optimum test reliability. Journal of Educational Measurement, 1975, 12, 109-113.

Ramos, R.A. & Stern, J. Item behavior associated with changes in the number of alternatives in multiple-choice items. Journal of Educational Measurement, 1973, 10, 305-310.

Tversky, A. On the optimal number of alternatives of a choice point. Journal of Mathematical Psychology, 1964, 1, 386-391.

JOB TAXONOMY/DUTY MODULE RESEARCH FOR OFFICER ASSIGNMENT DEFINITION

By

Arthur L. Korotkin Ph.D.

Presented at the
Eighteenth Annual Conference
of the
Military Testing Association
Gulf Shores, Alabama
October 1976

JOB TAXONOMY/DUTY MODULE RESEARCH FOR OFFICER ASSIGNMENT DEFINITION

By

Arthur L. Korotkin, Ph. D.

A. INTRODUCTION

The Army Research Institute for the Behavioral and Social Sciences (ARI) is conducting research to identify the qualitative personnel requirements of the Officer Corps in order to define an officer career progression system in support of the Officer Personnel Management System (OPMS). Sponsors for the research are the Chief, Research and Development, and the Deputy Chief of Staff, Personnel (DCSPER) of the Department of the Army. The research is designed to aid the Army and the individual officer in making career management decisions such as primary and alternate specialty designation, assignment, military and civilian education, and selection for promotion within OPMS. Specific objectives are:

> To develop a model career progression lattice, based on officer MOS, duty module, and skills analysis, delineating within-branch and cross-branch career development patterns leading to O-6 positions in the career progression programs of the OPMS.
>
> To relate experience and training requirements of the individual officer to the differential career assignment options identified through the development of the model career progression lattice.
>
> To develop and apply measures of interests, aptitudes, motivation, and performance for evaluation of the differential potential of the individual officer, and to relate these to the differential requirement of assignments, second specialty choice, training, and promotion.

The first of the research objectives, outlined above, is the development of an approach for describing jobs which more adequately defines the assignments in the Army officer force structure. A system is needed for describing and classifying jobs at a descriptive level which is detailed enough to provide the required information for career management purposes without being cumbersome and complicated to use. In addition, such a

CURRENT FOCUS

Military management information and decision systems

Oriented toward future years (5-7 years from now)

Capable of improving communications between organization planners and personnel management planners

Capable of facilitating decisions by a variety of military planning groups

system would provide a common language useful both in defining job requirements and personnel resources to fulfill these requirements. While such a system would have general applicability to the world of work, it is extremely important to the Army since its size, composition, deployment and hardware are continually changing necessitating constant adjustments in the training and utilization of personnel resources.

- Job descriptions currently available are not standardized and vary in level of detail from gross overall representations of the job to highly detailed descriptions of task elements comprising the job. The task elements involved tend to be too numerous and vary in their level of detail across the spectrum of jobs. At the other end of the scale, descriptions provided in the Military Occupational Specialty (MOS) system are quite general, providing only information for selection, assignment, training (other than with regard to a specific MOS), and the establishment of manning requirements.

B. THE DUTY MODULE CONCEPT

Although it was generally agreed that a new level of job description was necessary in order to be of use to both those dealing with resources and those dealing with requirements, the question still to be resolved was exactly how this generally useful "duty element" was to be designed.

Based on synthesis of available data, an approach to structuring the description of work activities evolved in which the following design criteria were applied: (a) the duty element must be meaningful and useful to requirement planners; (b) the duty element must be compatible with assignment practices in the field; and (c) the duty element must remain essentially the same even though the requirement may exist in a variety of assignments within the organization. It should be noted that if these criteria were met the duty element became meaningful to those concerned with training, the identification of assignments in which the duty element is a requirement, the identification of the level of training required, and the development of Programs of Instruction (POIs). The task cluster that resulted from the application of these design criteria were intended

NEED FOR GROUPING

NOT A MATTER OF "WHETHER" GROUPING IS DONE

GROUPING IS AUTOMATICALLY DONE ANYWAY
TRAINING PROGRAMS

PROFICIENCY TESTING "AREA" SCORES

THERE IS A NEED TO BE CONSISTENT AND
TO MINIMIZE ELEMENT AND CLUSTER
INCOMPATIBILITY

DUTY MODULE DEFINITION

A grouping of occupationally related tasks smaller than an MOS

"Modular" in the sense that it can serve as a "plug-in" unit for an assortment of MOS

CHARACTERISTICS OF DUTY MODULES

- DESIGNED IN TERMS OF A VARIETY OF MOS RATHER THAN JUST ONE
- ECONOMICAL IN THE SENSE THAT MANY DIFFERENT TASK INVENTORIES CAN BE DESIGNED WITH A SMALL SET OF DUTY MODULES
- UTILIZE DATA ALREADY COLLECTED BY MODB-1 ON ASSIGNMENT PRACTICES IN THE FIELD
- DESIGNED TO IMPROVE COMMUNICATIONS BETWEEN RESOURCE PLANNERS AND REQUIREMENT PLANNERS
- MEET APPARENT NEED FOR CONSISTENCY AT A LEVEL OF GENERALITY IN BETWEEN THE MOS AND THE TASK, AT AN OPTIMUM LEVEL OF DETAIL
- PERMIT MORE PRECISE MATCHING OF PERSONNEL TO JOB REQUIREMENTS AT A REASONABLE LEVEL OF DETAIL

to be self-contained, independent units of work, that would be modular in the sense that they could be used as "plug-in" units to a variety of different occupational specialties. They were named "Duty Modules."

The current procedure used for developing Duty Modules is a pragmatic one. It was shaped to a great extent by the means and resources which are available, convenient, and expeditious. First detailed job information is gathered on those positions deemed most important and representative by career managers in the Officer Personnel Directorate and which reflect the total (O-1 to O-6) grade spectrum. This job information is collected by a highly sophisticated group of senior Army Officers (retired) who have been trained as occupational analysts. The analysts then reduce the job data into task statements, and judgmentally cluster these statements into tentative Duty Modules ("Job Modules"). This clustering is guided by a set of criteria. Some of the more salient of these criteria are:

1. To be valid, the Duty Modules for any given position must be accurate and sufficient in describing the essential, truly significant, continuing work activity requirements of the position.

2. To be modular and useful, Duty Modules should be standardized, so as to apply in common across a number and variety of different positions and occupational specialties insofar as those positions actually have task clusters in common.

3. Each Duty Module should be a self-contained functional entity. It must not encompass, overlap, or depend on another Duty Module assigned to the same position.

4. A Duty Module should represent a distinctive, coherent, important part of the position. It may be important in terms either of criticality or proportion of time spent on it.

The Duty Modules are then validated (field verified) by subjecting them to review and critique by an independent sample of assignment incumbents. They are revised as necessary on the basis of these field verification data and added to the "Duty Module Catalogue."

Figure 1 is an example of an actual Duty Module (O-D-1) from the "Operations and Plans (Staff)" grouping. It shows the level of task

Figure 1

AIR Duty Module Survey Form (Rev1974)

Identification No._____

Code: 04-01

DUTY MODULE O-D-1 Performs operations staff functions in a general staff or other coordinating staff	(5) Direct	(4) Supervise	(3) Do and supervise	(2) Do	(1) Assist	(0) Not applicable
a Advise superior and others concerning operations matters.						
b. Prepare policy directives and SOP.						
c. Prepare and publish operation estimates and orders.						
d. Monitor execution of operations plans and orders and make changes as situation warrants.						
e. Recommend task organization, missions, and areas of operation.						
f. Organize and operate tactical operation center or operations element of command post.						
g. Determine operational readiness requirements and readiness status of unit.						
h. Recommend allocation of and authority for use of critical command resources such as replacements, special ammunition and aircraft.						
i. Coordinate overall security of command.						
j. Conduct or arrange operational unit readiness inspections and tests and take action to deal with problems.						
k. Prepare studies, reports, records, and correspondence pertaining to operations.						
l. Prepare and present operations briefings.						

	(0) Not applicable	(1) Little applicability	(2) Several of tasks	(3) Majority of tasks	(4) All of tasks
1. DO MODULE AND TASKS APPLY TO YOUR POSITION:					
a. In actual or simulated combat operations and support?					
b. In garrison and other than a?					

	(0) Not applicable	(1) 1–9%	(2) 10–29%	(3) 30–49%	(4) 50–69%	(5) 70–89%	(6) 90–100%
2. PERCENT OF TOTAL TIME SPENT ON THIS DUTY MODULE:							
a. In actual or simulated combat operations and support?							
b. In garrison and other than a?							

	(0) Not applicable	(1) Least critical	(2) Average	(3) Critical	(4) The most critical
3. RELATIVE CRITICALITY OF THIS PART (MODULE) TO ENTIRE JOB:					
a. In actual or simulated combat operations and support?					
b. In garrison and other than a?					

detail available in defining a Duty Module. In addition, as can be seen from the format (see Figure 1), when the module is identified within the requirements of a specific assignment, data are obtained on the level at which the incumbent performs the task, the percent of time spent on the Module, and the relative criticality of the Module to the entire job. The latter two are done for both combat and garrison conditions.

C. RESULTS

Highlights of this research program, and results relevant to officer personnel management are summarized below. Reports produced under these efforts are referenced.

1. The Duty Module structure has been completely developed and field verified for the Infantry, Quartermaster, Engineer, Ordnance, and and Armor career fields.

2. Duty Module requirements for significant grade and pertinent assignments in the Infantry and Quartermaster branches have been assembled and forwarded through the OPMS Task Group to the Infantry and Quartermaster schools for basic and advanced course staff review and evaluation for utility in POI development.

3. Selected job surveys have been conducted for a representative sampling (defined by OPD personnel) of assignments in the remainder of the OPMS specialties and additional Duty Modules developed as required; i.e., those needed to describe duties not defined by those Duty Modules already in the catalog. These preliminary "Job Modules" must be regarded as tentative until validated through field verification.

4. A matrix has been prepared showing the appropriateness of each Duty Module to each of the OPMS specialties. It shows which clusters of tasks (Duty Modules) are common to all OPMS Specialties and those that are unique to one specialty or a small number of specialties. Such a matrix is of value in analyzing officer training and in making

	AIR DEF ARTY	ARMOR	FIELD ARTY	INFANTRY	AUDIO-VISUAL INSTRUCT TE	COMBAT COMMUN. ELECT.	COMMUNICATIONS-EE	FIXED TELECOM. SYSTEM	COUNTERINTELLIGENCE HUMI	CRYPTOLOGY
A-1	X	X	X	X		X		X	X	X
A-2	X	X	X	X	X	X	X	X	X	X
A-3	X	X	X	X		X		X	X	X
A-4	X	X	X	X		X		X	X	X
A-5	X	X	X	X	X	X	X	X	X	X
A-6	X	X	X	X		X		X	X	X
A-7	X	X	X	X		X		X	X	X
A-8	X	X	X	X		X		X	X	X
A-9	X	X	X	X		X		X	X	X
A-10	X	X	X	X		X		X	X	X
A-11	X	X	X	X		X		X	X	X
B-1	X	X	X	X		X	X	X	X	X
B-2								X		
B-3									X	
B-4	X	X	X	X		X	X	X	X	X
B-5										
C-1										
C-2										
C-3										
C-4										
C-5										
C-6										
C-7										
C-8										
C-9										
D-1	X	X	X	X						
D-2	X	X	X	X						
D-3		X	X	X						
D-4			X							
D-5										
D-6	X	X	X	X						

Figure 2. A section of the Specialty X Duty Module Matrix

TABLE I

Indices of Commonality for Selected Pairs of OPMS Specialties*

Specialty	No. Unique Modules	Specialty	No. Unique Modules	No. Common Modules	Index
Infantry	26	Pers. Mgt.	2	16	.36
Maint. Mgt.	6	Gen. Troop Spt & Mat Mgt.	12	23	.56
Counterintell/HUMINT	1	Infantry	27	15	.35
Law Enforcement	20	R&D	4	5	.17
Armament Mat Mgt.	25	R&D	3	6	.18
Engineer	48	R&D	3	6	.11
Opns & Force Dev.	3	Tank & Ground Mob. Mat Mgt.	27	4	.12
Opns & Force Dev.	1	Engineer	48	6	.11
Opns & Force Dev.	1	Field Arty	34	6	.15
Procurement	4	Law Enforcement	17	8	.28
Procurement	4	Infantry	34	8	.17
Procurement	5	Finance	2	7	.50
Finance	6	Comptroller	6	3	.20
Infantry	20	Law Enforcement	3	22	.49

*CAUTION: These data are for illustrative purposes only since Duty Modules have not yet been completely developed for all of the OPMS Specialties.

determinations as to what type of training should be provided. It would also have impact on alternate specialty selection by showing the relationships among the 47 Specialties. It should be emphasized, of course, that the OPMS Specialty by Duty Module Matrix is not complete since not all of the jobs in the Army have been analyzed and not all of the Duty Modules have been fully verified. A section of the matrix is shown in Figure 2 for illustrative purposes.

5. A quantitative measure of Duty Module commonality across positions and specialties has been developed. Called the Index of Commonality it can be used to represent, with a single number, the ratio of Common Modules to total Modules within a pair of specialties or positions.

As a result of the efforts to date a total of 400 Job Schedules (Field job analyses and identification of selected job dimensions) have been completed. They represent, with varying degrees of coverage, the 47 OPMS Specialties. Grades have ranged from O1 to grade O6, with greatest coverage from O2 to O5. The selections of positions and incumbents were made with the assistance of the Career Branches of the Officer Personnel Directorate and covered most MOSs for troop assignments and all levels of staff assignments. In addition, thus far, 153 officer Duty Modules have been developed and most of them field tested in positions included in both TOE and TDA units ranging from platoon level to Hq. Dept. of Army Staff level.

The Matrix and Index of Commonality discussed above are of particular relevance to the OPMS career development effort. Presented in Table 1 are some examples of commonality indices for selected pairs of OPMS specialties. The utility of the data in Table I is illustrated by a more detailed expansion and discussion of the relationships between two specialties; e.g., Infantry and Law Enforcement (from Table I).

The Infantry Specialty contains 42 Modules and the Law Enforcement 25; of these, 22 Modules appeared in both. This gave them an Index of Commonality of .49. However, it should be noted that in going from Infantry to Law Enforcement, while there are 22 common modules, there are only three modules unique to Law Enforcement (not in Infantry).

From the above, a general hypothesis could be formulated as a basis for future detailed analysis:

"A senior Infantry officer, experienced in Infantry command and staff assignments could function well in the Law Enforcement specialty." It may be likely that the duties involved in both modules cc-1 and cc-2 may be learned as a part of on-the-job training and through familiarity with Law Enforcement activities at previous stations where the Infantry officer had served. Duty Module cc-4, however, includes tasks regarding rehabilitation and counseling. Since rehabilitation is a primary goal of military confinement, this type of work requires special training, probably in a civilian college, plus on-the-job training as a junior officer.

It can be seen that the breadth of knowledge, experience, and duties of an Infantry officer are notably different than a Law Enforcement specialty officer. It would appear that extensive additional training and experience would be necessary for a Law Enforcement specialist to function successfully in the breadth of Infantry assignments. The above analysis appears to fully support the OPMS mandate that denies Infantry specialty to officers that are not Infantry Branch.

Further research, using job analysis as a basis, may well result in a different designation of specialties since the primary skills of administration and management are reflected almost equally in all areas and are therefore common skills and knowledges. What may be revealed is that the differences in knowledge among certain specialties are not significant enough, particularly at senior levels, to merit separation as is currently reflected in DA Pamphlet 600-3.

E. PROJECTIONS

Future development and experimental application of the Duty Module concept needs to be accomplished. The next steps include the application of Duty Modules to:

1. Completion of the Duty Module Catalogue and development of an integrated and easily accessed data base of task and Duty Module information. This will involve the preparation of job schedules for all of the U.S. Army jobs not yet analyzed and the development and verification of additional Duty Modules as required.

2. Refinement of the tools and techniques for the development, analysis and application of Duty Modules. This includes demonstration and feasibility studies on:

a. The clarification of relationships between an individual's experience and training and his further assignment training and promotion alternatives.

b. The identification of training requirements and the identification of the most effective methods for fulfilling these requirements.

c. The facilitation of Second Specialty selection and development.

d. The specification of manning requirements and manpower forecasts.

e. The evaluation of both individual and unit performance.

3. The integration of methodology and results into a systematic program for application throughout the Army.

F. CURRENT RESEARCH

Current research efforts are focusing on research on the use of Duty Modules for:

1. providing feedback on training quality and/or deficiencies; and

2. the restructuring or "reengineering" of sets of jobs or positions using Duty Modules as the basic unit or building block.

ATTAINING EQUITABLE PERSONNEL ALLOCATIONS WITH SUPPLY AND DEMAND QUOTAS

Robert L. McMullen

US Army Research Institute Field Unit
Fort Rucker, Alabama 36362

Introduction

In the military enlistment promotion system, there is a unique situation that falls outside the bounds of the classical form of the transportation problem. The simplest form of progression is to advance from grade 1 to grade 9 while remaining in the same MOS. However, some MOS's terminate before reaching grade 9 and others do not begin at the grade 1 level. As a result personnel must be reclassified out of or into these MOS's. Furthermore, reclassification cannot occur arbitrarily from any given MOS to any other, but are instead constrained to allowable reclassifications specified by Army Regulations.

Sample Problem

Lets consider a hypothetical example of a promotion pattern of six MOS's between grade 6 and 7 (Figure 1). In this example, MOS A terminates at grade 6 and MOS C begins at grade 7. The arrows indicate allowable promotion paths. For example, personnel in MOS A must reclassify to MOS B or C in order to be promoted, and personnel in MOS's D and F must remain in their respective MOS's when promoted. The numbering of the cells is adopted for a convenience that will be apparent at a later point.

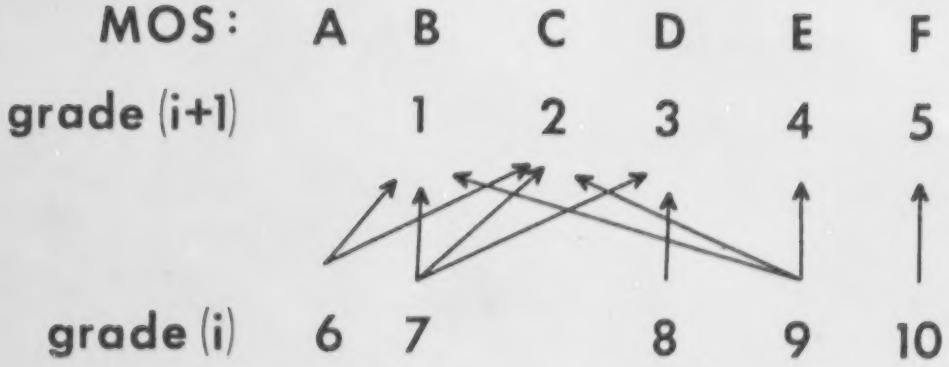


Figure 1

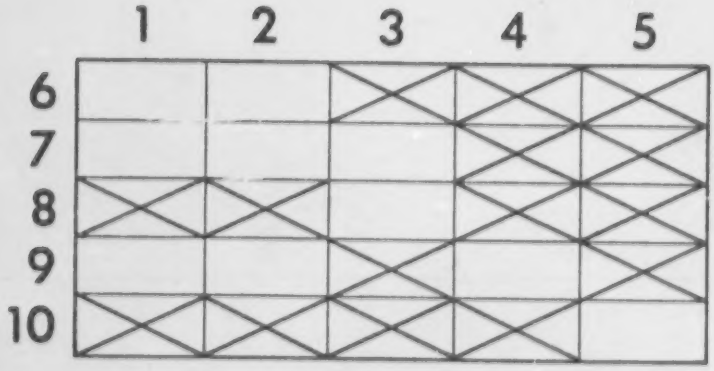


Figure 2

At the point when promotions are made, each of the MOS's at grade 6 will have a number of personnel that are promotable, i.e., supply, and each of the MOS's at grade 7 will have a number of vacancies or demands. The primary objective is to fill as many of the requirements as possible. The two secondary objectives are as follows: (1) assets that are not used would be as evenly distributed as possible among the suppling MOS's, within the constraints of the feeder pattern, (2) unfulfilled requirements should be distributed among the supplied MOS's as evenly as possible without sacrificing any promotions. In the real world, specific personnel policies may dictate a suboptimal solution.

Figure 1 can be redrawn in the standard transportation problem form as in Figure 2. The cells with X's in them indicate promotion paths that are not allowed. Transportation costs could be made by assigning zero costs to the empty cells and infinite costs to the X cells. At this point we hit our first objection with the classic transportation problem, which is that all assets will be assigned even if it results in an infinite cost. A second objection is that assets must equal requirements, which is rare in the real world. On the surface, these objections can be met by adding a dummy column to receive unused assets and a dummy row to absorb unmet requirements. A cost of 1 could be entered in these resulting cells to make them less desirable than actual assignments. The problem with this approach is that the values for the dummy assets and requirements are not known until the final solution for the original matrix is known.

Definitions

At this point it is desirable to state some of the characteristics of this type of assignment problem.

Supply Node - a source of personnel or items used to supply one or more demand nodes.

Demand Node - a sink which is fed by one or more supply nodes.

The number of supply and demand nodes may be unequal.

Assignment Matrix - a matrix that shows whether movement is allowed between a given pair of supply and demand nodes (Figure 2).

Open Path - a path from any supply or demand node to any other node.

Closed Path - a group of cells in the assignment matrix through which movement is allowed and forms a closed loop within the matrix.

'A' Matrix - this matrix is a table of the shortest path(s) between any pair of nodes, if a path exists. The 'A' matrix is square with the number of cells in a side equal to the sum of the number of supply and demand nodes in the assignment matrix.

Sample Solution

A sample problem using the assignment matrix from Figure 1 is shown in Figure 3, including the initial solution using the "northwest corner" method. It shows an unmet demand remaining in nodes 3 and 5, and unused supply in node 9. The problem is determining if there are paths from the node with excess supply to the nodes with the remaining demand, and that the excess supply can be moved along the paths. In this case, as

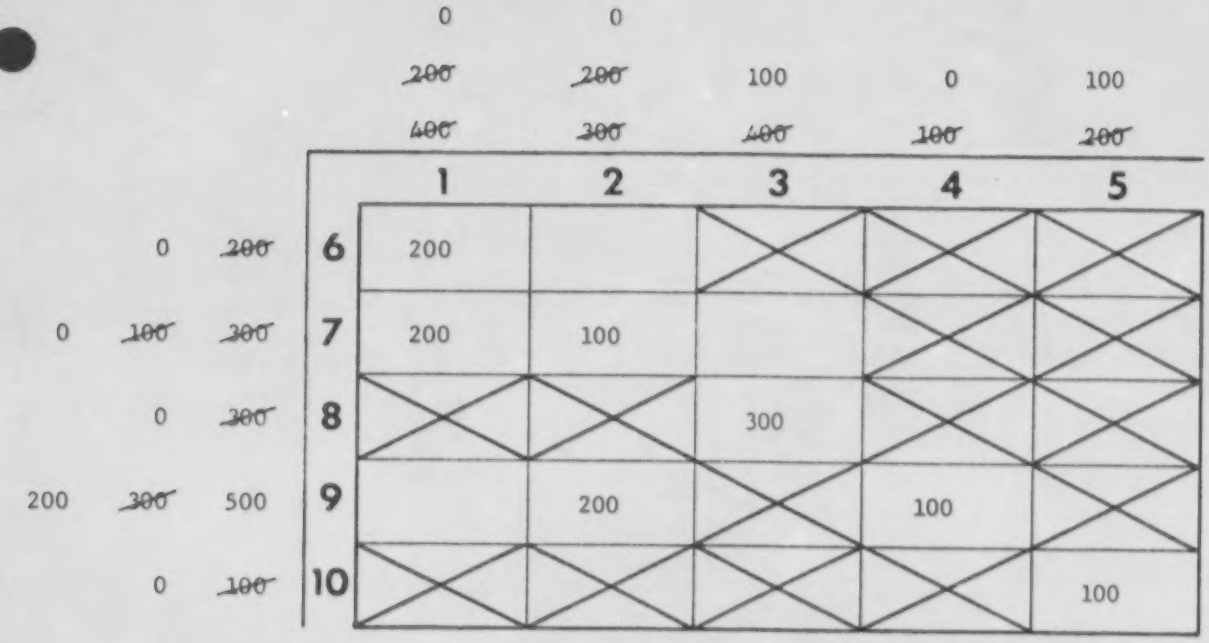


Figure 3

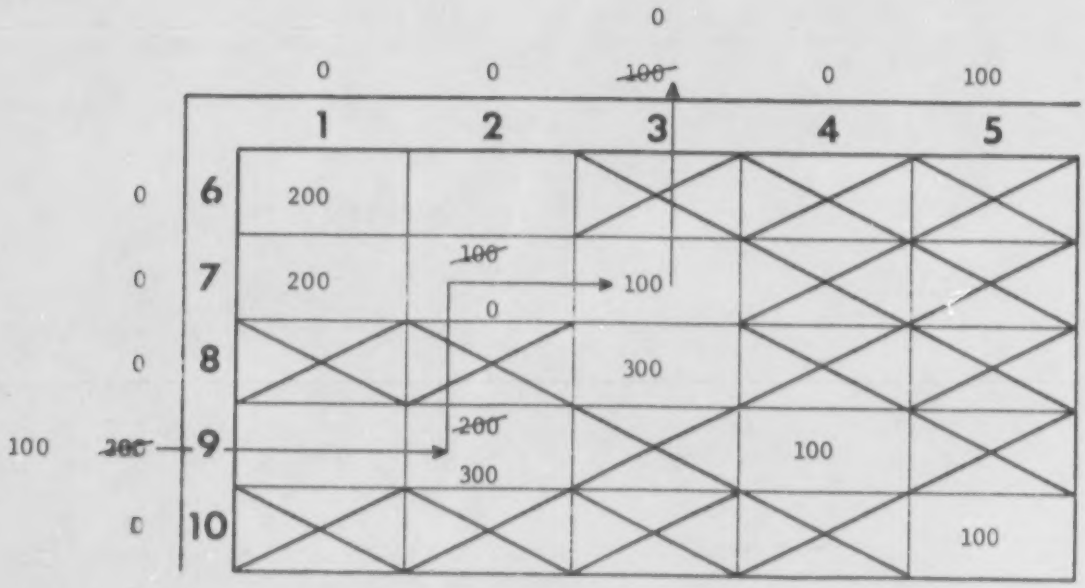


Figure 4

seen in Figure 4, one hundred of the supply can be moved to node 3. However, there are no paths between node 9 and node 5.

This solution is optimum with respect to the number of personnel that can be shifted. It may not be optimum from the personnel management viewpoint, i.e., one supply node used only 67% of its assets, while the others used 100%. A more desirable solution for management might be as shown in Figure 5 where the unused assets are spread proportionally among the supply nodes. The exception is node 5 which is not linked to any of the other supply nodes.

The key to the solution of this type of problem is determining the pathways between all pairs of nodes. Figure 6 shows only three of the possible paths between node 8 and node 2. Path number 1 is the shortest and the most efficient. The others require unnecessary shuffling of assets in columns 1 and 2, and rows 6, 7, and 9. A computer program that would identify all possible paths and select the shortest would be very inefficient for a matrix of any reasonable size.

The answer was to develop an algorithm that has the following characteristics: (1) explicitly identifies the shortest path from any supply or demand node to any other supply or demand node, (2) produces a table of the shortest paths, (3) identifies isolated subsets of nodes within the assignment matrix, if they exist, (4) locates closed loop paths within the matrix, (5) terminates automatically when the last shortest-path is found, (6) no constraint on size or shape of the assignment matrix, (7) is used only once for a given assignment matrix, and (8) uses matrix algebra. The

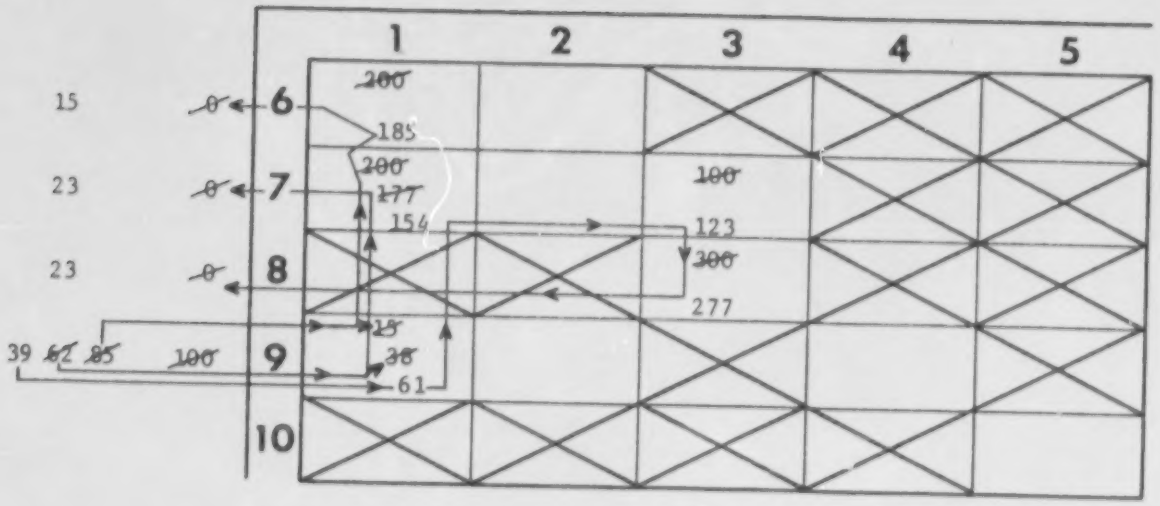


Figure 5

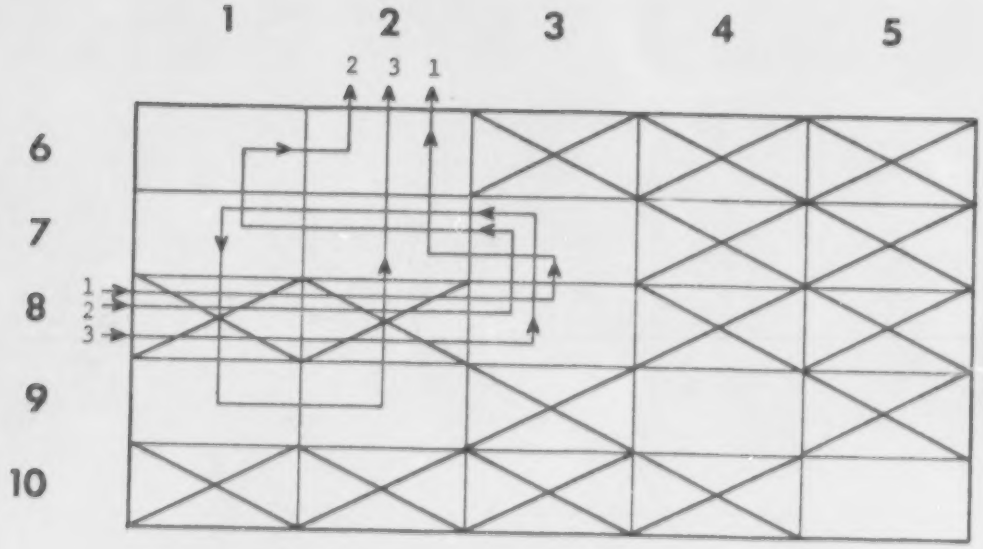


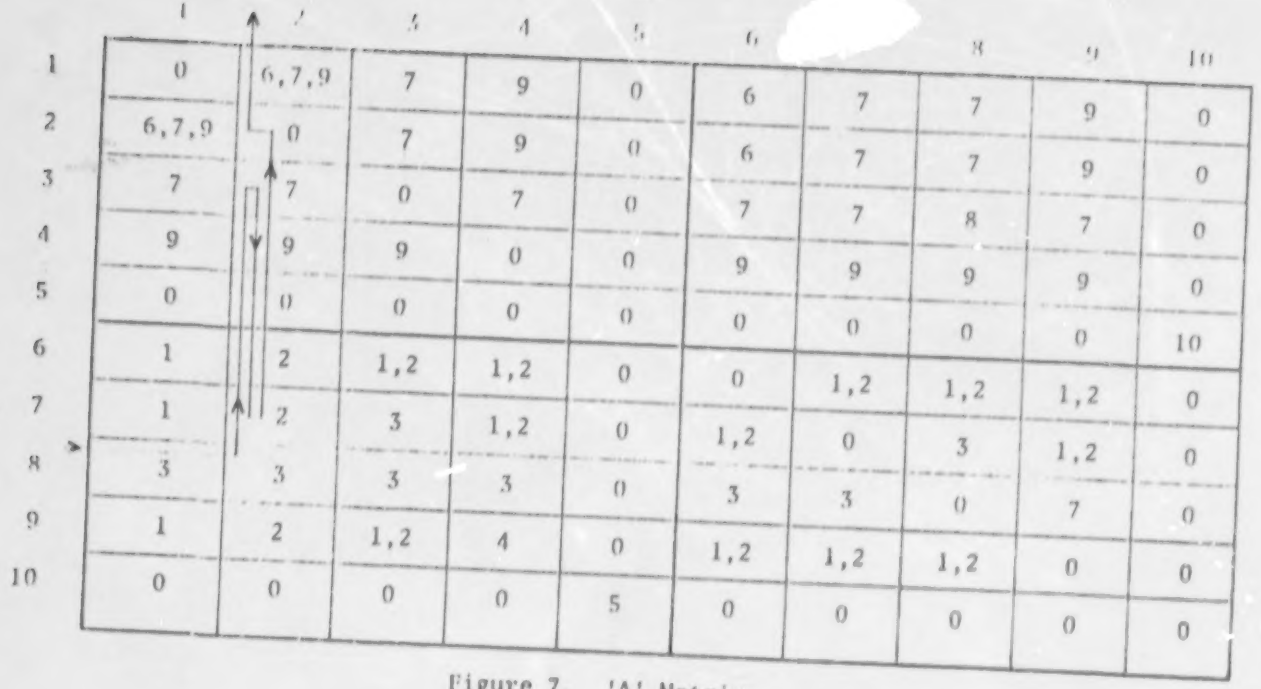
Figure 6

efficiency of the algorithm is due to the use of matrix algebra and the concept of identifying only the shortest path between any two nodes. It starts with the 2-step paths (in-pivot-out) then the 3-step paths, 4-step paths, etc. until termination.

The shortest-path table, or 'A' matrix, for the assignment matrix in Figure 2 is shown in Figure 7. The path from node 8 to node 2 (path 1, Figure 6) is also shown. Off diagonal zeros identify nodes that do not have any connecting links, i.e., isolated subsets. In the example, nodes 5 and 10 are connected to each other, but are isolated from the other nodes. Multiple cell entries in the 'A' matrix identify multiple shortest-path routes. These multiple entries also identify the closed loops within the assignment matrix.

Conclusion

A limitation in the past with personnel promotion problems using a sparse assignment matrix has been the difficulty in handling the complexities of the paths through the matrix. This complexity resulted in suboptimal solutions with respect to management objectives. The algorithm described above makes it possible to solve this type of problem more efficiently. At the same time, the solution can be more responsive to management's policies.



	1	2	3	4	5	6	7	8	9	10
1	0	6,7,9	7	9	0	6	7	7	9	0
2	6,7,9	0	7	9	0	6	7	7	9	0
3	7	7	0	7	0	7	7	8	7	0
4	9	9	9	0	0	9	9	9	9	0
5	0	0	0	0	0	0	0	0	0	10
6	1	2	1,2	1,2	0	0	1,2	1,2	1,2	0
7	1	2	3	1,2	0	1,2	0	3	1,2	0
8	3	3	3	3	0	3	3	0	7	0
9	1	2	1,2	4	0	1,2	1,2	1,2	0	0
10	0	0	0	0	5	0	0	0	0	0

Figure 7. 'A' Matrix.

Estimation of the Impartiality of Military Justice

Clark L. Hosmer, Ph.D.
Colonel USAF, Ret.

39 Longwood Dr.
Shalimar, Fl 32579

Presented at the annual conference of the Military Testing Association

Gulf State Park Resort, Gulf Shores, Alabama

October 1976

Estimation of the Impartiality of Military Justice

Clark L. Hosmer, Ph.D.
Colonel USAF, Ret.

My purpose is to estimate whether whites and blacks receive impartial justice in the military forces. My criterion is whether the probability is similar, within sampling errors as estimated by past records, of a white and a black receiving the same treatment.

In a system of perfect justice, no innocent person would be convicted and no quilty person would be acquitted. Also, two groups of people, drawn from a given population, would show only chance differences from each other in rates of crime, conviction, and severity of sentences. Whites and blacks are not drawn from the same population, however, so they may differ in their behavior and rate of commission of crime. Given the commission of a crime and entry into a system of justice perhaps it is reasonable to expect the investigation, release or trial, acquittal or conviction, sentencing, and serving time to be conducted impartially.

Juveniles and females differ from adult males in the eyes of the law. In addition, whether or not explicitly supported by legal codes, our system of justice appears to treat white collar offenders somewhat differently than blue collar and with still greater difference in treating indigent offenders. At the top of government, immunities are granted to officials and former President Nixon was treated with generous partiality.

With regard to race, however, beginning with the Reconstruction Amendments to the Constitution of 1865 and 1870 and then dramatically with President

Truman's July 1948 order to the Armed Forces to desegregate, the civil and the military systems of justice were not to treat white and black offenders differently under the law.

Civil Justice

Vontress/(1974) offers no data but describes partiality against blacks as customary. He concludes that a black is more likely than a white to be stopped on suspicion; if booked, more likely to be indicted; if indicted, more likely to be found guilty; if convicted, more likely to receive a severe sentence; and, when jailed, more likely to be required to serve his full sentence.

Simpson and Yinger (1973) report that a higher rate of blacks than whites who are charged with a crime are convicted and that more severe sentences are given blacks than whites. Data from the Federal Bureau of Prisons show that for prisoners committed by all courts to Federal prisons, the average sentence of whites was 43 months but 58.5, or 36% longer for nonwhites. The trend held when the data were controlled for age. The youngest group, 18 through 21 years, showed 57% longer average sentence for nonwhites and the next older group, 22 through 25 years, showed 67% longer average sentence for nonwhites.

Military Justice

In the Army, Navy, Marine Corps, and Air Force, the Uniform Code of Military Justice (UCMJ) provides a system unique to the Services and subject to command control that should help to insure impartiality of treatment of whites and blacks. The UCMJ spells out procedures for investigating, trying, sentencing,

and reviewing cases that range from those dealt with by the lowest level of so-called non-judicial treatment for minor violations through summary and special courts to the general courts that handle the most serious crimes.

Thomas, Thomas, and Ward (1974) studied the administration of non-judicial punishment aboard more than 70 ships of the Navy's fleets on the east and west coasts. They found a paradox. No documentary evidence of partiality against blacks was found. Furthermore, a statistically significant higher percent of blacks than percent of whites brought before the Executive Officer were dismissed by him without going to the Captain's Mast for final consideration and possible punishment (26% vs. 19%). On the other hand, the "overwhelming majority of blacks believed military justice favors whites."

How can one account for the belief in discrimination despite lack of evidence? First, at this lowest level of military justice, much of the preliminary action is not recorded. Before putting a man on formal report, a number of actions may be taken, including oral admonition, oral reprimand, variation in frequency and kind of work detail, and withdrawal of privileges. Comments by blacks on the Thomas questionnaires make vivid the alleged events of being put on report for minor or unfounded charges. Whether the events were factual cannot be determined by the study. What is suggested by the higher percent of blacks than percent of whites having their cases dismissed before facing the Captain of the ship is that in addition to formal evidence of impartiality of the Navy's administration of non-judicial punishment, the alleged partiality at informal levels may in part be remedied by the Executive Officers.

Turning to the full range of military tribunals in the Army, Navy, Marine Corps, and Air Force, the report of the Task Force on the Administration of Justice in the Armed Forces (1972) provides sufficient data to suggest a tentative answer to the question whether military justice is impartial with regard to race.

The report of the TaskForce repeatedly offers a red herring as a distractor from the question. The report cites the fact that a higher percent of blacks than their proportion in the military forces are involved with pretrial investigations, charges, trials, convictions, and severe sentences. But that might be due to a higher percent of blacks getting into trouble. It is not surprising that a higher percent of blacks may get into trouble if they feel that they have been put down and discriminated against. Such feelings would tend to make anyone fight the system. In any case, to suggest that a higher rate of blacks getting into trouble indicates partiality of the military system of justice wants further support to be acceptable.

The report of the Task Force shows three stages of military justice in which blacks do not fare as well as whites. The first is when a person is released from investigatory confinement with no further action. Of the whites, 33% were released as contrasted with 14% of the blacks, a statistically significant difference at the 1% level. The trend was the same when the cases were controlled for only those who had no prior record, those only who had non-judicial punishment, those only who had one or more prior courts martial, those only who were married, those only who were single, and those only who were absent without leave and had no other charge against them.

Thus, at the stage of those released without trial, the evidence suggests that military justice was partial toward whites and was harder on blacks.

The second stage that shows a significant difference in treatment is when a convicted person is sentenced. The whites' average sentence was 12 months as contrasted with an average sentence of 18 months for the blacks for a 50% longer term of confinement. The trend was the same when the cases were controlled for only those with no prior record, for only those who had prior non-judicial punishment, and for only those who had prior courts martial.

The third stage is when a person applies for an upgrading of the kind of discharge he received. Of the applicants, 31% of the whites and 20% of the blacks received relief of an upgraded discharge, significant at the 1% level. The trend is the same when the data are controlled for only those cases dealt with by the Army, Navy, and Air Force separately.

Research Needed

Stable estimates of the impartiality or differences in treating individuals would be useful to commanders, their staff judge advocates, and their chairmen of Advisory Council responsible for Social Action programs. The military reporting system has the capability of providing data on military justice as a flow process. Trends in what goes on within the system would help to illuminate critical stages and to show whether progress is being made.

Perhaps the Armed Forces can show the way to stengthening the hypothesis of equal justice under the law.

REFERENCES

Simpson, G. and Yinger, J. Racial and cultural minorities: an analysis of prejudice and discrimination. New York: Harper & Row, 1973.

Thomas, P. J., Thomas, E. D., and Ward, S. W. Perceptions of discrimination in non-judicial punishment. Navy Personnel Research and Development Center, San Diego, June 1972

United States Department of Defense. Report of the Task Force on the administration of Military Justice in the Armed Forces. 4 Vol. Nov 1972.

Ventress, C.E. Self-hatred in Americans of African descent. Unpublished paper presented at the American Psychological Association's annual convention. George Washington University, August 1974.

INDIRECT SAMPLING OF THE AFFECTIVE DOMAIN TO ASSESS "PETTY OFFICER QUALITIES"

by

Raymond D. Bland
LCDR, U. S. Coast Guard
Chief, Curriculum and Exam Development Division
U. S. Coast Guard Institute

The enlisted promotion system of the U. S. Coast Guard requires each candidate to meet a common set of standards in two separate evaluations. The first evaluation, at the local command level, determines by observation and performance testing that the candidate meets or exceeds job standards and may compete for promotion. The second evaluation is a norm-referenced multiple choice test given semi-annually to rank order qualified candidates from "greatly exceeds minimum standards" down to "barely meets minimum requirements." Ideally the test will be a reevaluation of a sample of the same factors evaluated at the local command.

Both evaluations are based on the Occupational Standards and the Military Standards of the U.S. Coast Guard Enlisted Qualification Manual. Items for Occupational Standards are selected from task statements in a Job Task Analysis. Military Standards are assessed inferrentially by defining affective traits of the "ideal" petty officer and selecting items from cognitive areas which are related to the affective traits.

The promotion system for enlisted members of the U.S. Coast Guard is a "best qualified" system which consists of three stages, each managed by different organizations. The first stage, managed at the local command level, is qualification for promotion. Each candidate must meet pre-

---

The proposals or assertions contained herein are the private ones of the writer and should not be construed as official or reflecting the views of the Commandant or the Coast Guard at large.

requisites of training, experience, performance testing, and evaluation by supervisors. The second stage, managed by the Coast Guard Institute at Oklahoma City, is semi-annual competition for a limited number of promotion opportunities. The competition consists of performance on a multiple choice exam administered to every candidate who has been certified at the local level as qualified for promotion to a particular rate. (The job specialty is a "rating" and each pay grade within a specialty is a "rate." "Radioman" is a rating, "Radioman Second Class" is a rate.) The third stage, managed by the Enlisted Personnel Division of the Office of Personnel at Coast Guard Headquarters, is the promotion of eligible candidates as vacancies occur. The purpose of the second stage in this three part process is to ensure that each promotion goes to the person who is best qualified for promotion at the time the vacancy occurs.

Both the first and second stages are dependent on the contents of the Enlisted Qualifications Manual. This manual comprises job descriptions for every rate in the Coast Guard. It tells the individual what he must be able to do in his current rate, tells him and his supervisor what the requirements are for the next higher rate (both knowledge and performance tests), and is the source document for the promotion exams developed by the Coast Guard Institute. There are two sets of qualification standards for each rate: Occupational Standards which are unique to the rating and Military Standards which are common to each pay grade across all ratings. Both sets of standards are considered in certifying each candidate as qualified for promotion in stage one and in writing the promotion exams in stage two.

The purpose of the exam in stage two is to discriminate among candidates who have all met the minimum requirements for promotion. By sampling job knowledge with a common test all candidates can be reevaluated against a common standard. The exam results, combined with points for experience and superior performance on the job, will produce a promotion list with candidates ordered from "greatly exceeds minimum requirements" to "barely meets minimum requirements." The exam will come closest to fulfilling its purpose if it is job relevant and has a high degree of job fidelity.

Achieving an acceptable degree of job fidelity is relatively easy on the Occupational Standards portion of the exam. Task statements from the "most-often-performed" portion of a Job Task Analysis are used to design the test items. Each statement used must also appear in the Enlisted Qualification Manual. This ensures that the item is based on a job task which the candidate is required to be able to perform and which incumbents perform frequently. The Occupational Standards also lend themselves to the design of separate sections consisting of several items on common subject matter. The Military Standards portion of the exam is much more difficult to deal with.

The Military Standards portion of the rate qualifications consists of a wide variety of duties which are related to levels of responsibility (i.e., pay grade) rather than to an occupational area (rating). Some, such as Training, Supervision, or Customs and Courtesies comprise routine tasks. Others, such as First Aid, Shore Patrol Duties, or Nuclear Warfare defense comprise contingency tasks which may be seldom or never performed, but which the individual may

be required to perform in extreme circumstances. All are standards which must be maintained for the Coast Guard to achieve its readiness status and for each individual to perform as a Petty Officer as well as a technician.

Each of the Military Standards has been met to varying degrees, by each candidate in stage one of the promotion process. As with the Occupational Standards, qualification in Military Standards is based on meeting or exceeding the standard without reference to how much the minimum level is exceeded. The Military Standards thus represent an additional dimension considered in the nomination for promotion. For this reason, they should be included in the competitive stage to achieve the goal of reevaluating the same factors originally considered. The contingency nature of many of the Military Standards makes job fidelity difficult to achieve since there is not a normal job task which the test item can simulate. Further, the wide variety of task areas (27 disparate areas) defies the design requirement of common subject matter in each section of the exam.

Prior to the development of the current set of exams (Fall 1976 to be administered in Spring 1977) the Military Standards portion of the exam consisted of one section with thirty items from each of the 27 topic areas. This did not achieve homogeneity of subject matter, but it did "cover the quals." It also avoided the issue of how to decide which topics could be dropped to achieve homogeneity and still achieve a reasonable sampling of each candidate's knowledge. During the Fall 1976 development we decided to face that issue.

A committee of Master Chief Petty Officers assigned as exam writers was appointed to resolve the problems of

homogeneity and job fidelity in the Military Standards portion of the exam. Early in their efforts they found that not only was the job task analysis not useful in contingency topic areas, but also they could not write meaningful task statements to fit many of the standards. Where statements could be written their job fidelity was very questionable. To resolve this impasse the committee was given a new mandate: Describe the process they had personally gone through in evaluating subordinates for promotion and from their separate descriptions find one or two common topic areas from Military Standards and discard the other areas. The goal was to eliminate less relevant areas so that further efforts would be concentrated on only a few areas of knowledge.

This brainstorming exercise produced a consensus on the nature and style of evaluation and a significant surprise which has taken us into a new area of exam design. All agreed that every evaluation they had been involved with and all they had ever been privy to was based on a review of the candidate's technical performance, both in the Occupational tasks of his rate and of the next higher rate. In addition, however, there was always a subjective opinion that "he will make a good ____ Class Petty Officer." No specific consideration of the Military Standards was involved in forming this opinion.

"He will make a good ____ Class Petty Officer" was expressed in a variety of ways, but always in terms of personality traits or habits; demonstration of affective traits, not cognitive knowledge. Technical competence on the other hand consisted of a recap of the candidate's cognitive knowledge and psychomotor skills in the Occupational Standards areas. What we were then faced with was writing an exam

consisting of a sampling of the cognitive domain for Occupational Standards and of the affective domain to parallel the non-occupational portion of the original evaluation. Military Standards seemed to be out of the picture. The problems of homogeneity and job fidelity were eclipsed by the necessity to sample the affective domain.

This new dilemma would be resolved if we could relate the affective traits to areas of cognitive knowledge, sample the cognitive areas and from the results of that sample make an inferrential assessment of the affective traits. This then would give us an exam which could assess the same areas which had been assessed, albeit subjectively, when the candidate was first considered for promotion. The rest of the development of the Military Standards portion of the exam was based on the hypothesis that

> Possession of some types of cognitive knowledge is demonstrated by an individual through the exhibition of positive personality traits which are observable; the absence of that knowledge leads to a lower level (or absence) of demonstration.

If the level of knowledge is related to the amount of demonstration, and the amount of demonstration has been assessed by the supervisor, then a common, comparative evaluation can be made, inferrentially, by testing the areas of related cognitive knowledge. High performance on the test of cognitive factors would imply high levels of demonstration of the traits of the affective domain. Low levels of performance would imply low levels of demonstration.

The committee of exam writers had developed a lengthy list of personality traits and habits in their brainstorming session. None of the factors was predominant, but the

majority were variations of two basic themes: concern for the physical well-being of self and others, and concern for the quality of interpersonal relationships on the job. There was consensus among the committee that subordinates they had known who were "good petty officers" were safety conscious and capable in first aid. Both Safety and First Aid are topic areas from the Military Standards and are close enough conceptually to comprise a common area of knowledge for exam design. Similarly, Customs and Courtesies, another topic area from the Manual, related to the traits which were variations of "concern for the quality of interpersonal relationships on the job."

This plunge into untried territory resulted in two new sections for the exams for E4, E5, and E6. (The exams for E7, E8, and E9 have a similar Military Standards section, but the standards are more structured and are mostly drawn from cognitive areas.) The old 27-topic section of low fidelity items across all the Military Standards has been replaced with two homogeneous sections which contain items with an acceptable level of job fidelity. The new sections provide an objective sampling of factors which were subjectively considered, but not ennunciated in the nomination process.

Basing the design of the Military Standards portion of the exam on the hypothesis of indirect relationships between the cognitive and affective domains allowed at least a rational choice of subject matter for the exam. Whether the relationship exists and to what extent it exists are questions which are neither resolved nor addressed in this project. What has been addressed and temporarily resolved is the question of how to design an exam for an area which is

amorphous, while remaining within the constraints of covering material that is in the job description and of designing homogeneous exam sections consisting of items with some degree of job fidelity.

ANOTHER LOOK AT PEER RATINGS

by

Major J. L. Mitchell & Mr. Angelo DeNisi

Department of Psychological Sciences
Purdue University
West Lafayette, Indiana

Peer ratings have a long history of being recommended as a credible source of performance information; that is, as a criterion measure (Wherry & Fryer, 1949; Hollander, 1954; Shenck, Watson & Hazel, 1973). Most of the basic texts on Industrial Psychology have recognized the potential of this type of rating. Campbell, Dunnette, Lawler, and Weick (1970) quote Barrett (1966) as concluding that peer ratings are potentially the most accurate judgements of employee job behavior. However, for all their potential, such ratings have an equally long history of being rejected by management as unsuitable for operational use; key executives have a keen distrust of what they view as popularity contests. Downey and Duffey (1975) call this concern "a bit of folklore" and noted that while there is a moderate degree of relationship between friendship and peer evaluations, this correlation is modest and can be controlled through the use of controlled distribution nominations. While their review of the previous literature suggested that the major portion of criterion variance was independent of friendship, nonetheless, as late as 1973, Shenck, Watson, and Hazel were concerned about the rejection of peer ratings as possible criteria of officer performance by Air Force Management. These authors urged that this management decision be reconsidered, so that the high potential of peer

ratings as performance criteria could be exploited in future Air Force research.

In much of the recent research, peer ratings have also been studied as to their possible use as predictors of future behavior, and several studies suggest that they have considerable utility for such a purpose. L.K. Waters and Carrie Wherry Waters (1970) used peer nominations as predictors of short-term sales performance with some success. Mayfield (1972) gathered peer nominations from 117 new life insurance agents in a three week training program and found significant predictive value for at least one variable against objective sales success criteria one year later. Amir, Kovorsky, and Sharon (1970) worked with promotions and officer training in the Israeli Defense Forces and obtained significant predictions through peer nominations with correlations generally around .60 and.70. Their best predictions were found to be highly related (r = .79) to selection for and success in officer training. They had the advantage of working with a very structured organizational environment but their study deserves serious consideration because of their success with a very large sample (n = 3897). Amir, Kovorsky, and Sharon concluded that peer nominations would make a marked contribution to the predictive validity of conventional psychological screening measures.

Downey and Duffey (1975) have recently conducted a rather extensive review of the various approaches to peer ratings and so we do not need to detail those here. They concluded that peer ratings were highly reliable and, although subject to some influences (such as friendship, familiarity, sex, race, etc), were of great potential utility. Downey and Duffey observed that peer evaluations have been an exceptionally fruitful tool in both research and application and they suggest that the various problems encountered with the use of peer ratings are soluable. They highly recommend the use of

peer ratings as a powerful tool for evaluating complex human behavior.

In this paper, we would like to take a somewhat different view. We believe that there is some evidence to suggest that management has been correct to reject the use of peer ratings in operational performance evaluation systems, and we think there is now some reason to believe that such ratings do become distorted in operational use. However, while such peer ratings may not be adequate for operational use as criteria, we feel that they are still useful both as possible predictors of success and as research criteria (as, for example, as a performance criteria for test validation). Further, we would like to assert that there are other potentially valuable uses for peer ratings which may well be even more significant that their use as performance criteria. For example, peer ratings may be of very great utility in building realistic and meaningful motivational systems through their use as one source of performance feedback to the individual worker.

Before we discuss this suggestion, we need to elaborate on our belief that such ratings are somewhat distorted in operational use. Although the evidence supporting the validity of peer ratings is rather impressive (c.f., Kraut, 1975; Roadman, 1964; Weitz, 1958; Mayfield, 1970 among others), little research has been conducted relative to the question of whether the ratings of all peers are equally valid. It would seem quite reasonable to believe that, in any given situation, some peers would be in a better position to make ratings than others. Certain variables worthy of consideration are more obvious than others (such as familiarity), but there are also a number of variables growing out of the evaluation procedure itself that should be investigated.

In their review of peer rating methodology, Downey and Duffey (1975) included some of the variables that have already been investigated. The

first which they discussed was length of association. It is clear that group members must be in contact with each other for some reasonable period of time before valid evaluations can be made. Although the research has been fairly consistent in showing that valid evaluations can be made after as little as 3 weeks (for example, see Hollander, 1965), this period will vary when new members are introduced into an intact group. As Mayfield (1975) points out, this period of time will be greater when a new person enters an already formed group, and Downey (1975) has shown that an individual not well known to other members of the group will be rated near the middle of the groups distribution. Although there is no direct data on the subject, we would also expect that the ratings made by such a new member of the other members of the group would be somewhat distorted until he gets to know the other members better. It is evident, then, that the length of association between group members will have an influence on the way peers are rated.

Another issue which must be raised, is whether association in itself is enough, or if the nature of the association is also important. Freeberg (1969) found that peer evaluations were more highly related to a performance criterion when the interaction between peers occurs in an area relevant to the performance dimension being rated. Therefore, it would seem that not only the length of association is relevant as a variable in peer ratings, but also, for peer ratings to be used optimally, raters should have had a reasonable association with other group members where the relationship has touched on areas relevant to the performance dimensions to be evaluated.

When we discuss length of association and relevant interaction, we must also assume that friendships will be developing as people work together and get to know each other. As noted earlier, one of the strongest arguments used by management against the operational use of peer ratings is that these ratings will be influenced by friendship. The data has usually indicated that although

friendship does result in "higher" peer ratings, it does not greatly influence the validity of peer ratings (Hollander, 1956, 1965; Waters & Waters, 1970). This seeming contradiction is usually attributed to the fact that these studies employed nomination techniques which allowed the rater to give good ratings to non-friends as well. Also, while most studies have concluded that friendship is not a major contributer to criterion variance (Downey and Duffey, 1975), the fact remains that it does usually make some contribution and thus must be considered as one additional potential bias in peer ratings.

An issue closely related to friendship involves the formation of subgroups among members - particularly racial subgroups. There is some evidence that racial differences exist in peer ratings (de Jung & Kaplan, 1969), although it appears that such effects may be confounded with friendship. Relative subgroup size may also be an issue since Schmidt and Johnson (1973) report no race effects when there are equal numbers of black and whites being rated.

Sex is also an issue. Downey & Duffey (1975) observed that in a small sample of Army officers, females scored lower than males on ratings on the part of both males and females. Interestingly, females' self ratings were not related to either male or female evaluations, but males' self ratings were related to these evaluations. Thus, it is clear that ratings may be given differentially to males and females, although we do not yet understand the specific causes of these differences.

All of these variables (length of relevance of association, friendship, race, and sex) can be expected to have some influence on peer ratings and, as such, they must be considered when we evaluate the utility of such ratings. However, even though these variables are enough to make the situation complex and difficult to understand, they are not the only varialbes which must be considered in assessing the value of peer evaluations.

Schneider and Bayeroff (1953) studied peer ratings in General Staff School classes and noted that those people who received the highest ratings also gave the most valid ratings. There seems to be some contamination present in this study since all the data was collected at one session and the people who were giving the peer ratings were also in a position to influence the careers of their peers. In any event, it does seem that there would be differences among people in rating their peers. It also appears reasonable to conclude that the people who are "best" on the job are also the ones best able to rate peers' performance on that job.

Another interesting issue is pointed up by Kaufman and Johnson (1974). These authors note that negative ratings contribute very little to the predictive validity of peer ratings, and they conclude that only positive ratings should be used. Doll & Longo (1962), however, had already found that if they identified those ratings which were low for "antisocial" reasons, the validity of these "low" ratings increased. This might suggest that the ideal rater, in a peer rating situation, is one who is well acquainted with the peers he is ~~going to~~ rate, but is able to separate his evaluations of the peers on social and personal bases from his evaluation of their actual performance on the job or their potential for future success. This sounds very similar to Fiedler's categorization of Low LPC (task-oriented) leaders (Fiedler, 1967), and perhaps this concept would be useful in identifying those individuals who are able to give the most accurate peer ratings.

A final perspective on peer ratings relates to the rating process itself. A recent small field study (DeNisi & Baum, Note 1) looked at the effects of a person's receiving a low ratings by his peers on his subsequent ratings of the peers. This field study was conducted in an operational setting since the peer ratings received by an individual were to constitute on quarter of the individual's semester grade in an upper level (mixed graduate and undergraduat

business school course. The results of this study tend to support the notion that people retaliate if they receive low ratings. A comparison of the mean rating given on subsequent rating cycles by such individuals indicated a general tendency for these ratings to be lower than the mean of their ratings on the initial phase, while this was not generally the case for people who received high ratings. There was a general trend for ratings to decrease over time.

Further, this study pointed up a number of differences among sub-groups of raters in terms of their discrimination in ratings. Graduate students, including several military officers, were better able to discriminate in their ratings as indicated by closer agreement with a criterion (instructor's mean rating) and had significantly greater variance across ratees. They also had greater "loyalty" to other subgroup members; the military officers were the most cohesive subgrouping present and showed a strong tendency to rate other military people higher than anyone else; they were slightly more critical than other graduate students in rating everyone else.

This is an important study since it is one of the few situations where peer ratings have been evaluated in actual operational use. Further, this study lends credence to management's rejection of the use of peer ratings in that it demonstrates that peer ratings do become distorted in operational use, with those who receive low ratings tending to "retaliate" against their peers. Thus, in addition to the possible influences of friendship, familarity, race, and sex, potential users of peer ratings as criteria of performance must also be conscious of the dynamic interaction between the ratings received and subsequent rating behavior.

The study also illustrates that, when a sub-group is cohesive enough, they will not only rate other group members with unfair leniency, but will also tend to rate non-members with unfair strictness. It is interesting to note that all the military personnel except one, were in the Army, which would

foster even greater cohesiveness. This suggests that especially in the military, where espirit de corps is so important, peer ratings will tend to be seriously distorted by sub-group membership. One has only to think of the possibility of pilots and navigators rating each other in peer ratings to see the implications of such sub-group membership on ratings. Or even finer subgroupings could be used, such as bomber, transport, and fighter pilots. The potential number of subgroups is quite high and the interaction possibilities become very complex indeed.

The DeNisi and Baum study also suggests a possbile solution for the problem of selecting those people most capable of making valid peer ratings. Although the graduate students, and particularly the military officers, also tended to receive the highest peer ratings (although the military people still rated the other military people higher than anyone else did), this is probably due to superior experience in making presentations of the type used in the class. This experience has also allowed them to get a better idea of what a "good" presentation should be, and so enabled them to be more discriminating in their ratings. This implies that, if we are seeking to identify and utilize the most reliable and accurate set of peer raters, we should probably seek those individuals who have had more experience in the area being rated, since these people will have a better idea of what constitutes "good" and "poor" performance and will this be better able to discriminate in their ratings.

Downey and Duffey (1975) have suggested that it is not feasible nor cost-effective to collect individual ratings from every individual in large groups. They note that the number of ratings to be collected equals N^2 where N is the number in the group. We suggest that many of the difficulties and the high cost associated with the collection of peer ratings can better be controlled by the identification of a selected subset of peer raters; that is, those individuals in a given speciality who have proven their own capabilities

and who have had experience in the specific performance areas being rated. While the possible influence of other variables still need to be studied in some detail, we suggest that this sub-group, once identified, could provide more reliable and valid ratings than would be possible through using ratings by all members of the group.

Given that we are able to identify individuals who will provide relevant and reliable peer ratings, what is the optimum use of such ratings? Certainly such ratings can be used where needed as predictors of selection and success, as in the cases of the work of Amir, Kovorsky, and Sharon (1970) and Waters and Waters (1970). And we see no objection to the use of such ratings as criteria of performance in those research situations where many of the operational distortions would not be present, such as in test validation.

However, we suggest that the use of peer ratings as an operational criterion of individual performance is contraindicated; the various influences on such ratings in operational situations are just too diverse and are not well enough understood or controlled.

This does not mean, however, that we feel that there are no useful applications for such ratings. Rather, we believe that there are a number of potential uses. Once such use which we would like to discuss here would be in the area of performance feedback to individual workers to enhance their understanding of their own performance (knowledge of results).

One of the most consistent findings in the research literature is the lack of agreement between individual workers and their supervisors on the nature and the quality of work accomplished by the individual. For example, O'Reilly (1973) in a study of 57 pairs of supervisors and workers, found only one pair of the 57 where there was complete agreement between the supervisor and worker on what the basic tasks of the worker's job were. Borman (1974) observes that this lack of agreement may stem from the differences in perspec-

tive, which derive in good part from the differing positions in the organizational structure (that is, different organizational levels). Zedeck, Imparato, Krausz, and Oleno (1974) reached essentially the same conclusion from their study of behaviorally-based rating scales. This and related research suggests that perhaps the most useful performance information would be that which is generated at the same organizational level as the ratee. Mitchell (1975) has suggested that the performance rating policy of peers is more like that of an individual's self rating than it is the rating policy of the supervisor.

Further, Bassett and Meyer (1968) note that performance information which is received from a supervisor is apt to be met with some hostility on the part of the worker, which interfers with any motivation for improvement in performance. This finding implies that information on performance which is received from sources other than the supervisor might be more acceptable to the worker, and hence more useful to the individual. Thus, the research literature would seem to indirectly support the idea that peer ratings may have greater potential as performance feedback information than do the performance appraisals of supervisors.

Feedback, in this context of performance ratings, can be defined as the flow of information back to the individual concerning his own performance and his relationship to his job situation. It would appear that meaningful feedback is a very critical need before an individual can perceive that his performance can and in many cases should be improved. Yet, it also appears likely that how he receives the information (and from whom and in what form) can make a considerable difference in how the individual accepts, evaluates, and uses this performance information (Pritchard and Mitchell - Note 2).

In many organizations, feedback has become institutionalized into performance counseling sessions or as formal performance report ratings. Typically, these types of performance assessments occur at regular intervals of

six months or a year. Such appraisals often tend to be tied to a formal reward system involving merit pay increases or promotions. In such situations, the feedback information available in such sessions is far overshadowed by reward allocation information. Such performance reviews provide only very generalized and nonspecific information. As McGregor pointed out almost two decades ago, performance appraisals seldom contain information critical of the employee's performance (McGregor, 1957). Kay, Meyer, French (1965) in their work at General Electric, have concluded that knowledge of results (feedback) is important for the worker but cannot be accomplished effectively in a single annual interview. They recommend separating pay or promotion reviews from performance coundselling, and suggest that supervisors should give more frequent informal feedback including both praise and information on recommended improvements. Bassett and Meyer (1972) have suggested that feedback be based on the worker's self report which would be brought to the supervisor and discussed. These authors have utilized an operational system based on this concept and report that it resulted in improved motivation and hence better performance and job satisfaction.

An alternative approach to feedback improvement might be through the use of peer ratings to provide performance information to the individual. When considered in the light of our earlier suggestion that such ratings would be more economical and reliable if they were derived from a selected subset of the peer group, this becomes a very practical performance feedback system. With such selected-peer ratings (derived from the best people in any specialty), information could be obtained economically on the relative standing and specific performance problems of each individual. This type of information could be presented to each individual directly, without supervisors knowing the ratings. This would provide useful information to the worker as to how he stands with his peers but should be less threatening, since the supervisor

would not be involved in the process. Such objective information would provide the individual with an opportunity to see himself as others see him, and to be alerted to the need for change and improvement. We cannot be sure how the individual would react to this kind of feedback. Future research is needed to find an optimum procedure for such feedback so as to maximize the probability of a positive motivational response.

This suggestion is put forward as an alternative to using peer ratings only as a criteria of performance or as a predictor of future performance in a traditional personnel system. As noted earlier, there is some question as to how such ratings would fare in operational use. The present suggestion is somewhat outside the traditional personnel system, yet would serve a useful purpose by providing the individual with performance information he is not currently receiving.

Reference Notes

1. DeNisi, A. & Baum, J., Some contextual problems with the use of peer ratings in an operational setting. Unpublished manuscript, Purdue, 1976.

2. Pritchard, R. D. and Mitchell, J. L., Toward more effective performance. Manuscript, Department of Psychological Sciences, Purdue University, 1976.

Bibliography

Amir, Y., Kovorsky, Y., and Sharon, S., Peer nominations as a predictor of of multistage promotions in a ramified organization. Journal of Applied Psychology, 1970, 54, 462-469.

Barrett, R. Performance rating. Chicago: Science Research Associates, 1966.

Bassett, G. A., and Meyer, H. H., Performance appraisal based on self review. Personnel Psychology, 1968, 21, 421-430.

Borman, W. C., The rating of individuals in organization; an alternative approach. Organizational Behavior and Human Performance, 1974, 12, 105-124.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., and Weick, K. E., Managerial behavior, performance, and effectiveness. New York: McGraw-Hill, 1970.

deJung, J. E., and Kaplan, H., Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. Journal of Applied Psychology, 1969, 53, 518-524.

Doll, R. E., and Longo, A. A., Improving the predictive effectiveness of peer ratings. Personnel Psychology, 1962, 15, 215-220

Downey, R. G. Associate evaluation: nominations vs. ratings. U. S. Army Research Institute for the Behavioral and Social Sciences, 1974.

Downey, R. G., and Duffey, P. J., Various methodological approaches to peer evaluations. U. S. Army Research Institute for the Behavioral and Social Sciences. Prepared for the twenty-first Conference on the Design of Experiments in Army Research, Development and Testing, Washington, D.C., 1975.

Fiedler, F. E., A theory of leadership effectiveness. New York: McGraw-Hill, 1967.

Freeberg, N. E., Relevance of rater-ratee acquaintance in the validity and reliability of ratings, Journal of Applied Psychology, 1969, 53, 518-524.

Hollander, E. P. Buddy ratings: military research and industrial implications. Personnel Psychology, 1954, 7, 385-393.

Hollander, E. P. The friendship factor in peer nominations. Personnel Psychology, 1956, 9, 435-447

Hollander, E. P. Validity of peer nominations on predicting a distant performance criterion. Journal of Applied Psychology, 1965, 49, 434-438.

Kaufman, G. F., and Johnson, J. C., Scaling peer ratings: and examination of the differential validities of positive and negative nominations. Journal of Applied Psychology, 1974, 59, 302-306.

Kay, E., Meyer, H. M., and French, J. R. P., Effects of threat in a performance appraisal interview, Jouranl of Applied Psychology, 1965, 49, 311-318.

Klimoski, R. J., and London, M. Role of the rater in performance evaluation. Journal of Applied Psychology, 1974, 59, 445-451.

Kraut, A. I., Prediction of managerail success by peer and training staff ratings. Journal of Applied Psychology, 1975, 60, 14-19.

Mayfield, E. C., Management selection: buddy nominations revisited. Personnel Psychology, 1970, 23, 377-391.

Mayfield, E. C., Value of peer nominations in predicting life insurance sales performance. Journal of Applied Psychology, 1972, 56, 319-323.

Mayfield, E. C., Peer nominations in the life insurance industry. IN: R. G. Downey and F. F. Medlan (Chair) Peer Ratings: Beyond Validation Studies, Symposium presented at the Meeting of the American Psychological Association, Chicago, 1975.

McGregor, Douglas., An uneasy look at performance appraisal. Harvard Business Review, 1957, 35 (3), 89-94.

Mitchell, Jimmy L., Capturing performance rating policies from multitriat-multirater data. Procedings of the 17th Annual Conference of the Military Testing Assoication, Indianapolis, Indiana, 15-19 Sept 1975: pp. 377-386.

O'Reilly, A. P. Skill requirements: supervisor-subordinate conflict. Personnel Psychology, 1973, 26, 75-80.

Roadman, H. E., An industrial use of peer ratings. Journal of Applied Psychology, 1964, 48, 211-214.

Shenck, F., Watson, T. W., and Hazel, J T., Relationship between personality traits and officer performance and retention criteria. Technical Report AFHRL-TR-73-4. Personnel Research Division, AFHRL, Lackland AFB, Texas, May, 1973.

Schmidt, F. L., and Johnson, R. H., Effect of race on peer ratings in an industrial setting. Journal of Applied Psychology, 1973, 57, 237-241.

Schneider, D.E., and Bayeroff, A.G., The relationship between rater characteristics and validity of ratings. Journal of Applied Psychology, 1953, 31, 278-280.

Waters, L. K., and Waters, C. W., Peer nominations as predictors of short term sales performance. Journal of Applied Psychology, 1970, 54, 42-44.

Weitz, J. Selecting supervisors with peer ratings. Personnel Psychology, 1958, 11, 25-35.

Wherry, R. J., and Fryer, D. H., Buddy Ratings: popularity contest or leadership criteria? Personnel Psychology, 1949, 2, 147-159.

Zedeck, S., Imparato, N., Krausz, M., and Oleno, T., Development of behaviorally anchored rating scales as a function of organizational levels. Journal of Applied Psychology, 1974, 59, 249-252.

The Chaplain Description Scale: A Peer Evaluation Without Tears

by

Chaplain (LTC) John S. Snyder
Chaplain (LTC) Duie R. Jernigan
Chaplain (MAJ) Cecil D. Lewis

Presented at the Military Testing Association Annual Conference
18 - 22 October 1976

Gulf Shore, Alabama

The Chaplain Description Scale: A Peer Evaluation Without Tears

INTRODUCTION

In 1972 the Army Chaplain School began a new system of peer assessment among its Officer Advanced Course students. It was necessary; the traditional method of peer ratings used there was immensely unpopular. Rank ordering their fellow students seemed like saying that one was a better person than another. Students didn't want to say that, particularly when it meant that some would be at the bottom. Moreover, many students who seemingly should not be worried were terribly anxious that they would rank low. Whether the rating was one general ranking of all-round personal effectiveness or of a few traits such as leadership, contribution to group effort, or attitude, the system was bitterly perceived as a "popularity contest."

Nevertheless, the Chief of Chaplains needed an assessment of interpersonal strengths—information not gained from academic averages. How effective a leader is this student? How well does he work with other chaplains? Faculty advisor evaluations were not adequate alone; often they were affected as much by the rater's predispositions as by the rater's deportment. In order to lessen the effect of individual differences in perception, a sufficiently large number of raters would be needed, and this necessitated student participation. In order to lessen student frustration with rank ordering or naming as outstanding a limited number of their classmates, we needed a descriptive, rather than an ordering scale. Finally, in order to furnish to the chaplain career branch office the information necessary for its planning purposes, discriminating evaluations on a number of traits had to be obtained for the official report.

The Chaplain Description Scale (CDS) is a 20-item, semantic differential rating scheme which is completed on each student by all the members of his learning group. In the Chaplain School the students do much of their work in 11-13 person groups, to which they belong for the duration of the course. The CDS can be completed by a student on all the members of his group in less than 30 minutes. Each student's profile is computed from the average of the ratings made in the 20 items of the scale, which is then compared with the group average on each item. Interpretation is made through profile sheets and explanation by experienced counselors.

DEVELOPMENT

The areas of concern, as indicated earlier, came from the Army Office of the Chief of Chaplains, which needed information to plan career patterns and assignments. Existing inventories were examined which produce profiles based on points on a scale between contrasting poles of personality factors. These included Kostick's Perception and Preference Inventory, the Personal Orientation Inventory, and the Sixteen Personality Factor Questionnaire. Our scale differs from these in that participants rate each other and that the traits are assessed directly, rather than with a large number of indirect questions.

The following elements of interpersonal effectiveness comprise the CDS:

Teamwork

Gave little of himself to team effectiveness (minimal collaboration).	Gave freely of himself to team effectiveness (maximal collaboration).

We ask the students here to rate each other on the degree of cooperation, involvement, and interest which was shown in their group work. We see this factor as predictive of an individual's ability to contribute effectively in a staff or directorate situation.

One of the requirements on the Army Student Academic Evaluation Report is to rate Contribution to Group Work. We determine that rating from this item on the CDS.

Oral Expression

Seemed to have trouble "thinking on his feet."	Was quick, articulate, and expressive "on his feet."

A relatively high score indicates that, in extemporaneous situations, the individual speaks clearly, smoothly, and convincingly. This item also is the source used at the Chaplain School for one of the required ratings on the academic report.

Creativity

Displayed a low degree of creativity (not an initiator of ideas).	Displayed a high degree of creativity (an initiator of ideas).

In the small group seeting students are expected to deal with case studies, projects, or complex problems. Successful group work in the course, as will also be the case in future staff assignments, demands sufficient creative ideas. How much a student has provided pertinent, innovative, and stimulating ideas to the group process is requested in this item.

Empathy

Appeared insensitive to the needs and feelings of others (not "person-oriented").	Was sensitive to the needs and feelings of others ("person-oriented").

How well the chaplain works with others, how effective as a leader he is, and how successfully he counsels are all related to his responsiveness to others' feelings. His degree of empathy will become quite clear to his peers over a period of several weeks in the small group.

Attention Seeking

Displayed little desire to be noticed.

Consistently drew attention to himself.

One of two items in the CDS which are not clearly positive at one end of the scale, Attention Seeking is an attribute which provides useful descriptive information about an individual's style. The flamboyant or assertive individual may be very desirable in some kinds of assignments and less so in other kinds. Interpreted in conjunction with other elements of the CDS, Attention Seeking fills out the picture of an individual's characteristic mode of interpersonal action.

Openness

Was closed, cautious, and defensive in his dealings with others.

Was open, candid, and non-threatened in his dealings with others.

One of the indicators of effective participation in a group, Openness indicates the degree to which an individual has been willing to share his reactions and feelings. An absence of hidden agendas, withdrawal, and delayed responses will result in a high rating. Openness is not only an important quality in cooperative staff work, but also it has been demonstrated to add to counseling effectiveness.

Leadership

Was generally ineffective and unsure of himself as a leader.

Was a highly effective, self-confident leader.

Leadership skills, as perceived by his peers in a learning group, are extremely important in the consideration of a chaplain for future supervisor and staff positions. Such a peer evaluation of this trait is considered highly valid because leadership is exerted without the benefit of differing functions or higher status.

Self-Direction

Showed a low degree of self-direction.

Showed a high degree of self-direction.

Students observe their fellows to have more or less initiative in starting projects or stating opinions. The person with a high degree of this trait appears to be inner-directed, able to work independently, and is not easily swayed by the group.

Competitiveness

Demonstrated an excessive degree of competitiveness.

Demonstrated a healthy degree of competitiveness.

Most of the officer students are ambitious and are aware of the importance of performing well in the course. Some betray their tendency to compete excessively by refusing to cooperate with the group or seeking to impress an instructor unduly. Such behavior in a training environment can be expected to continue in a regular assignment.

Colleagueship

Seemed to experience difficulty in relating to chaplains of other faiths.

Related effectively to chaplains of other faiths.

"Cooperation without compromise" is a frequently voiced aim of the Chaplain Branch. How fully a chaplain can relate to chaplains holding quite different beliefs will have a profound effect on his ability to work on a staff with other clergymen.

Dominance

Showed little interest in controlling the decisions and actions of others.

Displayed a strong desire to control the decisions and actions of others.

Like Attention Seeking, this item has no positive pole. The two traits generally go together, i.e., a high score in one is matched by a high score in the other. If a high Dominance score by a student is accompanied by a high Leadership score, he is seen as a strong, assertive leader. On the other hand, low Leadership and high Dominance indicate that he has been trying, but not succeeding, to influence his group.

Authority Orientation

Related poorly to persons in authority (apprehensive, tense, anxious).	Related well to persons in authority (poised, self-assured, composed).

A student's peers observe him relating to Staff and Faculty members and compare his behavior to his customary style. Examples of poor Authority Orientation would be an obsequious attitude on one hand and a belligerent, fighting-the-problem attitude on the other.

Perspective

Consistently focused on minor issues (missed the "big picture").	Characteristically perceived the larger issues (saw the "big picture").

Group members must work out among themselves their priorities, courses of action, and a number of issues where common sense and judgment are important. It is apparent that good judgment in such a setting would correlate positively with good professional judgment on the job.

Administrative Proficiency

Showed low proficiency in working with administrative details.	Showed high proficiency in working with administrative details.

The care with which a student completes his portion of a project or assignment, including minor or related aspects, will determine how he is rated. This sort of attention to detail in order to tie up all loose ends is the same kind of trait necessary for thoroughness in administrative work in the field.

Social Integration

Mixed poorly with people (detached, aloof, stayed to self).	Mixed well with people (warm, affable, outgoing).

Certain individuals in an average group of chaplains will stand out as "person-oriented". They will be ranked high in Social Integration and will also tend to score high in such items as Empathy, Openness, Colleagueship, and Facilitativeness. A cluster of high ratings in these areas coupled with relatively low scores in other areas usually occurs with an individual who favors a pastoral ministry rather than a staff or supervisory position.

Task Orientation

Seemed unwilling to put in long hours to complete tasks.	Displayed a willingness to put in long hours to complete tasks.

Student learning groups frequently are strained to complete objectives and must devote extra time to them. How willing individuals are to go past the scheduled class time will be reflected here. This trait correlates positively with Teamwork.

Advocacy

Was reluctant to present differing views to superiors (avoided confrontation).	Was a frank advocate of ideas to superiors (constructively confrontive).

Related to but only one aspect of Authority Orientation, this item helps to identify those who are more outspoken and frank in carrying their concerns to persons in authority. Certain chaplain positions require such a trait more than do other positions.

Organizational Ability

Demonstrated low organizational ability in the accomplishment of tasks.	Demonstrated high organizational ability in the accomplishment of tasks.

This can be interpreted either in terms of organizing people or of working out efficient, systematic approaches to complex problems. A high score usually is matched by a high rating in Leadership.

Facilitativeness

Tended to hinder learning in others.	Tended to facilitate learning in others

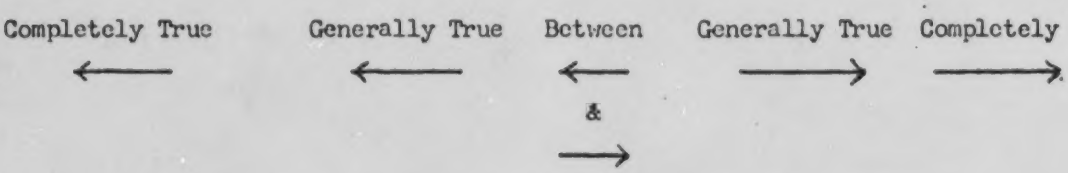
Of key importance in the group process of learning used at the Chaplain School, Facilitativeness denotes helping, supportive efforts to fellow members. As mentioned above, it is generally high if other person-oriented traits are rated high.

Poise

Was characteristically "uptight" in his dealings with others.	Was characteristically "at ease" in his dealings with others.

Poise is, of course, an asset for any person. It is especially important for a chaplain in his ministry and staff roles. This trait generally correlates positively with Social Integration.

The twenty items just described are printed on an IBM mark-sense form. Between the contrasting statements are five spaces. The rater marks one of them according to his selection of one most nearly reflecting his judgment of the person he is describing. They are as follows:



In our experience students have been able to choose comfortably from these five possibilities.

ADMINISTRATION

Officer students, including those in the Basic as well as the Advanced Course, take the CDS twice, the first half-way through the course and the second near the end of their training. They are told at the beginning of training that they will later be requested to participate in it. By means of a computer program, the CDS forms are preprinted with the names of the students, both raters and ratees. Thus, Chaplain Smith will receive a packet of 12 forms. The first will have printed on it, "Packet for Chaplain Smith; Person being rated: Chaplain Jones." Succeeding forms will have the names of the other members of Chaplain Smith's group. Also, the rated individual's Social Security number and group identification information are pre-slugged on the form so that the CDS can be scored by machine.

It is important for students to understand that the ratings will be compared only within their own learning group. It is not the absolute average score which matters; rather it is the individual's average score in relation to the average rating of all in his group. We say to the students that in our experience there has always been a normal range of effectiveness within every learning group. We expect that they have seen clear differences among their colleagues during the past several weeks.

They should, therefore, use the scale to indicate differences. To those who might be tempted to help all in their group with high ratings, we point out that scores can only be considered high in terms of the average of that group. If all members of a group should give a maximum score to everybody in the group, there would be nobody above the average and therefore nothing could be noted as a special trait for anybody in the group. In the years of our involvement with the CDS, students, once they have realized this fact, have tended to use the middle to upper range of the scale, and clear discriminations have been apparent.

OBTAINING RESULTS

A computer program furnishes a printout of each student's average ratings for the 20 factors. They are expressed numerically, since the computer reads the ratings on a scale from 1 to 5. Also on the printout is the group mean for each factor. This is the mean of all the group members' average ratings on each trait. Obviously this will vary from group to group, and the class average may well be quite different from one group's mean on a particular factor. We are interested, however, only in the deviation of a student's rating from the average of his own group. To make this explicit, we have the computer furnish the standard deviation of the distribution of individual average scores for that factor. Finally, we are given the Z score of the student for each factor. Unusually high Z scores (approaching +1.00 or higher) are clear indications that an individual's peers see him as exceptionally high in those areas. A printout is analyzed by an experienced staff member, who can see the significance of high scores involving natural clusters of traits. One individual may not have a single trait as high as +1.00 Z score, but does show +.50 or more on several related factors. Such a result would cause the evaluator to give the student a higher-than-average rating in that general area.

All reports are made in narrative form, and the printouts are destroyed after an appropriate period of time by the Chaplain School. To write the official academic reports, the evaluator uses the printouts from the two administrations of the CDS. He uses phrases such as the following: "As seen by his fellow students, Chaplain Smith demonstrated superior ability to offer creative and significant ideas in problem-solving tasks." Because our practice is to identify strengths more than weaknesses, we tend to say a good deal more about the individual who scores high in several areas than we do about the chaplain who stands out in none or very few.

INDIVIDUAL FEEDBACK

Up to this point, the only reason discussed for our use of the CDS at the Chaplain School has been its input for official reports. The value of the CDS for individual self-understanding and growth, however, is at least as important as information for reports. It is stimulating to most people to see how they are perceived by others, and almost invariably our students have welcomed the opportunity to receive such data. It must be done with care, however. Cold figures may be misunderstood, and failure to excel may be dispiriting. Our evaluators are skilled counselors, and appointments are made to give an interview to every student individually after each of the two administrations of the CDS. A person therefore has the opportunity to discover his weak areas halfway through the course and then work to increase his effectiveness before the second administration of the rating.

In order to portray a student's rating more clearly than Z scores appear to most people, a CDS Profile Sheet is utilized. The 20 factors are listed with the same contrasting statements and the computer prints on the form the point at which his average rating falls. The evaluator draws a line to complete the profile and show graphically which traits appear most characteristic of him. With an acetate stencil, the evaluator then pinpoints the group mean of the 20 factors, connects them with a line, and points out where the student is above and below the group mean.

After the second CDS has been taken, the results are discussed with the student together with any changes from the first profile.

During the administration of the CDS, each student is asked to complete a form on himself. The heading reads, "Packet for Chaplain Smith. Person being rated: Chaplain Smith." This rating does not contribute to his average score. We explain that the self rating is in no way used for an official report. Its value is in the feedback session, when we can help each student compare his self rating with the profile given him by his fellow group members. This self appraisal appears on a profile sheet in the same way as his average ratings. His self rating is a series of points directly under the 5 options. A more jagged line than his average profile is normal. We help him see whether there is general congruence or not, e.g., are the highs of his self rating the same factors as the highs of his average rating? If not, he may be helped to realize a blind spot in his self-perception, or he may realize the need to work harder at demonstrating his "real" style of behavior.

VALIDITY AND RELIABILITY

In the three years of its use, the CDS has met generally with very positive reactions. Students usually find no surprises in their profiles; they confirm their impressions of how others see them. Similarly, the faculty evaluations of students have, on the whole, corroborated the CDS rati s. Systematic tests of validity, however, have not yet been made.

Three retest reliability tests have been made, each with samples of about 12 individuals. Advanced Course students rated each other in a retest after one month. The average correlation coefficient (Spearman's rho) was .69. Among a group of Basic Course students, a retest after one week resulted in an average correlation coefficient of .76. A group of Staff and Faculty members, retested after one day, produced a correlation coefficient of .85.

The results of all 20 factors for each group are shown in the following table:

FACTOR	CDS TEST-RETEST GROUP & INTERVAL Advanced	Basic	Staff & Faculty
	1 month	1 week	1 Day
TEAM WORK	.83	.81	.81
ORAL EXPRESSION	.90	.65	.99
CREATIVITY	.95	.82	.93
EMPATHY	.48	.76	.94
ATTENTION SEEKING	.90	.79	.94
OPENNESS	.91	.85	.90
LEADERSHIP	.88	.73	.94
SELF-DIRECTION	.74	.68	.82
COMPETITIVENESS	-.05	.53	.88
COLLEAGUESHIP	.44	.79	.93
DOMINANCE	.89	.70	.81
AUTHORITY ORIENT	.65	.88	.78
PERSPECTIVE	.87	.84	.87
ADMIN PROFICIENCY	.96	.66	.91
SOCIAL INTEGRATION	.48	.64	.76
TASK ORIENTATION	.39	.75	.64
ADVOCACY	.66	.88	.81
ORGANIZATIONAL ABILITY	.89	.85	.89
FACILITATIVENESS	.23	.72	.81
POISE	.86	.85	.71

DISCUSSION

The CDS has contributed significantly to the evaluation of students at the Chaplain School, both for official reports and for aiding the students in developing interpersonal skills. It has worked well particularly in the small group process of our courses, because students have a great deal of opportunity to become familiar with the behavior of the 11 or 12 other students. Self-paced instruction would not be as suited for the CDS.

Validity studies are still needed. The ratings give clear indications of skills in areas which should result in successful performance in certain chaplain positions. Ratings of performance by supervisors of graduates in the areas covered by the CDS would be of value in determining predictive validity. Possibly a comparison of Officer Efficiency Reports with the CDS ratings will provide useful information.

A limited effort has been made to carry the CDS scores a step further by determining the relative values of the factors in various chaplain positions. This effort has not been completed. It will be necessary first to identify key chaplain positions and to verify the degree to which the factors contribute to superior performance. If this proves successful, we can provide more direct information about aptitude for various career options. The Chaplain Branch staff people who work in career management will have more certain data for their critical decisions in assignments and training. Finally, the students can be better counseled by our institution in their plans for career goals and self-development.

The CDS can be adapted for use in other schools and organizations. Items need to be developed which are appropriate for the field in which the participants will work. For the system to operate successfully, they will need to have worked closely enough together and for a sufficiently long period of time, so that each member can easily rate all the members of his group. A computer program is recommended because of the large amount of data. For only 13 participants there will be 144 ratings on each

of the 20 factors, or 2880 scores to use in the calculation of means, standard deviations, and Z scores. With computer assistance the CDS has proved to be an efficient instrument for securing an enormous amount of significant information. With these procedures and the appropriate modifications, we believe that the CDS can be an effective instrument for evaluation in many other institutions.

RELIABILITY OF ASSOCIATE RATINGS OF PERFORMANCE POTENTIAL BY ARMY AVIATORS

Robert F. Eastman and Robert L. McMullen

Presented at the 18th Conference
of the Military Testing Association
18-22 October 1976

Reliability of Associate Ratings of Performance Potential by Army Aviators

BACKGROUND

In response to a TRADOC request, the Army Research Institute Fort Rucker Field Unit has undertaken a program to determine the attributes which make some attack pilots outstanding performers in combat. The program which was proposed consists of the following three inter-related tasks: (1) development of an attack pilot profile from survey data of proven performers; (2) development of a rating form to evaluate potential attack pilots; and (3) evaluation and assessment of AH-1 trainees using the findings of tasks 1 and 2.

Currently no systematic selection of candidates for AH-1 transition exists. Many trainees do not want to be attack pilots but are assigned to fill requirements because they are due for reassignment. If unit commanders and personnel had more information, a better fit of aviators to transition training might be possible. This research is part of task 2 and was done to determine the reliability of unit-level ratings of potential AH-1 candidates.

OBJECTIVES

The primary objectives of this study were to evaluate the inter-rater reliability of a rating form designed to select AH-1 transition candidates and to study rater acceptance of rating forms.

It was hypothesized that COBRA pilots in cavalry and assault units would demonstrate high inter-rater reliability when rating non-AH-1 qualified aviators in their units as on potential for AH-1 transition and gunship pilot duties. A variable of interest was the effect of length of rater-ratee acquaintance on inter-rater reliability (Freeberg, 1969).

Structured interviews of more than 50 attack pilots revealed that about 70% felt that they could predict other pilots who were potential "ACEs" (i.e., attack pilots who perform exceptionally well in combat) without observing them under combat conditions. This preliminary research suggested that a majority of attack pilots would favor use of rating forms to evaluate candidates for AH-1 transition in contrast to a general lack of acceptance of ratings to determine leadership potential in Army schools (Medland, Yates and Downey, 1974; Downey, 1975).

SAMPLE

RATEES - The ratees were UH-1 and UH-1/CH-58 qualified aviators in six troops of the 6th ACCB, one troop of the 2d Armor Division, and one troop of the 1st Cavalry. For some units all eligible ratees were used. For those troops which had more than 12 eligible ratees, a random procedure was used to eliminate individuals and reduce the number of ratees to 12. The modal number of ratees was 12 (see Table 1). Because of anonymity requirements imposed by unit commanders, no analysis of the backgrounds of ratees was possible.

RATERS - The raters were all available AH-1 qualified aviators in the eight aviation units. There was considerable variation in the number of raters available within each troop (see Table 1). Because of field duty assignments many AH-1 rated aviators were not available for participation in this study. However, no systematic basis for non-availability which would affect this study was apparent.

PROCEDURE

Several troop commanders were unwilling to have their personnel participate in this study unless careful procedures were used to insure the anonymity of raters and ratees. This requirement was satisfied as follows: The raters in each troop were scheduled to arrive at the classroom as a group. Each rater within a group was given a packet with a randomly preassigned two-digit number as he entered the classroom. When the entire troop arrived, the list of names of ratees for the troop was placed on the blackboard and a set of three-digit numbers were randomly assigned to each ratee. The raters were told that the set of numbers on the board corresponded to three-digit numbers appearing on the evaluation forms in their packets. It was clearly stated that no previous connection between names and numbers had existed. Raters were then instructed to consider the entire list of individuals before following the written instructions on the rating forms. Immediately after each troop completed their ratings the names and numbers on the blackboard were erased. No records of ratees' numbers were kept.

RATING SCALE

The rating scale was modeled after a general format used at several Army schools[1] and is designed to encourage discrimination among ratees

[1]For example the RANGER Evaluation Report; FB (RD) Form 2.

NUMBER OF RATES AND RATEES IN EACH UNIT

UNIT #	TROOP DESIGNATION	NUMBER OF RATERS	NUMBER OF RATEES
1	B 7/17	11	12
2	C 4/9	6	12
3	C 7/17	10	11
4	A 7/17	10	12
5	D 1/9	6	12
6	D 2/1	6	13
7	HHT 4/9	3	13
8	A 4/9	7	12
	TOTALS	59	97

[illegible] of [illegible] characteristics for gunship/attack pilots. The attack pilot characteristics were identified during structured interviews of attack pilots with combat experience. On the evaluation form the rater is instructed to consider the attack pilot characteristics and assign the ratee a numerical rank (i.e., 1-25 in a typical group of 25 pilots) representing potential for success as an attack pilot. The rater is also provided space for a 2-3 sentence word picture to justify the numerical rating assigned. Additional information is recorded on where the rating was conducted and the type and duration of the relationship between rater and ratee. Detailed instructions also spell out the restrictions to be followed when rating a group of AH-1 candidates (see Appendix A).

RATING FORM EVALUATION

The rating form evaluation was designed to ask AH-1 qualified raters the following: (1) characteristics to add/delete to the list on the rating form, (2) if ratings should be used and who should do the rating, and (3) open-ended questions about why they do or don't favor ratings and potential operational problems of using ratings.

RESULTS AND DISCUSSION

Coefficients of concordance were computed for each unit to determine the degree of agreement among raters. The results shown in Table 2 indicate that inter-rater agreement was very high ($p < .01$) for 6 of the 8 units studied. Although statistical significance was not quite achieved in units 6 and 7 ($p < .02$), the number of raters involved was small. The overall results indicate that raters were responding to essentially the same characteristics when rating the candidates in their units. This does not mean that the observed orderings were necessarily correct, or based on the "true" criteria for a successful gunship pilot; however, the location of the attack pilot characteristics on the rating form itself was designed to make the rating criteria explicit and reduce the influence of irrelevant variables.

The results of the Rating Form Evaluation are presented in Appendix B. The percentage of the sample (if over 5%) who responded to an alternative or who made an open-ended comment is shown on a copy of the rating form itself. The most important result was that 58% of the respondees felt ratings should be used in selecting candidates for COBRA transition training. However, many raters indicated reservations about the validity of ratings because of the influence of personal prejudices and friendships (20%) and the adequacy of ratings to assess candidates effectively (15%). To the question "who do you think should rate AH-1 candidates", only 20% of the sample responded "commanding and supervisory personnel", while 58% indicated either AH-1 qualified pilots (29%) or other pilots (29%) in the unit.

SIGNIFICANCE OF OBSERVED COEFFICIENTS
OF CONCORDANCE FOR ALL UNITS

UNIT #	W	CHI-SQUARE	P <
1	.381	46.049	.001
2	.541	35.706	.001
3	.502	50.200	.001
4	.559	61.490	.001
5	.378	24.948	.01
6	.342	24.624	.02
7	.724	26.064	.02
8	.349	26.870	.01

Another variable which is important in peer rating research is how well the rater knows the individual being rated. The measure of association recorded in this study was the length of time the rater knew the rated officer. The relationship between median duration of rater-ratee acquaintance within each unit and the statistical reliability of inter-rater agreement is shown in Table 3. The data in Table 3 suggests that inter-rater reliability, in these units, was positively related to the duration of association between raters and ratees.

CONCLUSIONS

AH-1 qualified aviators demonstrate a high degree of inter-rater reliability when rating AH-1 candidates in their units. Statistical reliability was higher within those units where raters and ratees had been acquainted longer.

The majority (58%) of AH-1 qualified raters sampled feel that ratings by fellow aviators should be used in selecting candidates for COBRA transition.

Many raters (20%) had reservations about the use of ratings because of the influence of friendship and other personal biases. Another group felt ratings could not (8%) or might not (7%) be effective in identifying potentially successful AH-1 candidates.

It is apparent that further research is required to determine if the high degree of agreement among raters is valid with respect to an external performance criterion. A research project currently underway will attempt to validate unit-level ratings of AH-1 students as a predictor of transition course grades.

REFERENCES

Downey, R. G. Associate evaluations: Improving field acceptance (Research Memorandum 75-5). July 1975.

Freeberg, N. E. Relevance of rater-ratee acquaintance in the validity and reliability of ratings. Journal of Applied Psychology, 1969, 53, 518-524.

Medland, F. F., Yates, L. G., and Downey, R. G. Associate ratings of senior officer potential (Research Problem Review 74-2). June 1974.

RELATIONSHIP BETWEEN INTER-RATER RELIABILITY AND DEGREE OF RATER-RATEE ACQUAINTANCE WITHIN UNITS

Unit #	Median Number of Months of Rater-Ratee Acquaintance	Significance Level of Inter-Rater Agreement
1	18	$p < .001$
2	12	$p < .001$
3	12	$p < .001$
4	15	$p < .001$
5	6	$p < .01$
6	9	$p < .02$
7	10	$p < .02$
8	6	$p < .01$

Page

Appendix

A. ATTACK PILOT CANDIDATE EVALUATION FORM 9

B. RATING FORM EVALUATION 10

ATTACK PILOT CANDIDATE EVALUATION FORM

Instructions:
1. Evaluate this man in your unit/class in terms of your estimate of his potential ability to become a successful gunship/attack pilot. Determine where you think he would rank in a typical group of 25 pilots (number 1 the highest ranking, 25 the lowest ranking). Consider the ATTACK PILOT CHARACTERISTICS below prior to rating each man. Consider the entire group you are asked to evaluate and the following restrictions before beginning. (a) No more than two individuals may be placed in 1-5 column. (b) no two individuals will be assigned the same rating number. Do not rate yourself.
2. Under REMARKS, write a 2-3 sentence word picture to justify the numerical rating you assigned. State briefly the characteristics (desirable or undesirable) of this man that impressed you most.
3. Your ratings will remain anonymous. The packet you picked up has an ID number only to insure that you followed the restrictions when rating.

EVALUATED INDIVIDUAL'S NAME (Last, first)

DATE
DAY / MONTH / YEAR

ATTACK PILOT CHARACTERISTICS

TACTICAL KNOWLEDGE	AGGRESSIVENESS	CONFIDENCE
PHYSICAL STAMINA	SELF-DISCIPLINE	TEAMWORK
TIMELINESS OF ACTION	DRIVE	INITIATIVE
COMBAT INSTINCT	EFFECTIVE MAP USE	DEPENDABILITY

EVALUATED PERSON'S PRESENT LOCATION (Circle one)	IERW	UNIT	TRANSITION TRAINING
RELATIONSHIP TO EVALUATED PERSON (Circle one)	CO	IP	OTHER

STANDING WITHIN A 25-MAN GROUP (Circle one)

1	6	11	16	21
2	7	12	17	22
3	8	13	18	23
4	9	14	19	24
5	10	15	20	25

REMARKS:

HOW LONG HAVE YOU KNOWN THE INDIVIDUAL? _____

RATER ID #

USAAVNC/ART Fm 1793, 31 Mar 76

1. What characteristics would you add to the list on the rating form?

	% Responding
Desire	17%
Common Sense	17%
Mental Stamina	7%
Maturity	5%

2. What characteristics would you delete from the list on the rating form?

	% Responding
Combat Instinct	14%
Physical Stamina	8%

3. Who do you think should rate AH-1 candidates?

AH-1 qualified pilots only	All other pilots in their unit	Commanding and supervisory personnel only	Instructor pilots only	Senior aviators only
		% Responding		
(29%)	(29%)	(20%)	(14%)	(8%)

4. What problems, if any, can you foresee in using a rating form like the Attack Pilot Candidate Evaluation form at the unit level? Use the reverse side if more space is needed.

	% Responding
Personal prejudices and friendships will invalidate it	20%
Form can't judge qualifications	15%

5. Do you feel that ratings by fellow aviators should be used in selecting candidates for COBRA transition?

	YES	NO
% Responding	(58%)	(42%)

Briefly indicate in the space below why you feel ratings should/should not be used. Use the reverse side if more space is needed.

Should	% Responding	Should Not	% Responding
Fellow aviators more aware of abilities	12%	Ratings are biased	8%
		Ratings are no good	8%
		Ratings may not be any good	7%

October 1976

TOTAL TESTING IN COAST GUARD AVIATION TRAINING

BY

Harry R. Anderson

and

H. Kingsley Povenmire

Department of Transportation

U.S. Coast Guard

AVIATION TRAINING CENTER

Mobile, Alabama 36608

The opinions expressed herein are those of the authors and should not be construed as official or reflecting the views of the Commandant or the Coast Guard at large.

Harry R. Anderson

H. Kingsley Povenmire
October 1976

BACKGROUND

Coast Guard rotary wing aviators annually attend a one week Proficiency Course at the Aviation Training Center in Mobile, Alabama. The purpose of this training is to review instrument and emergency procedures and renew their instrument rating. The guiding principle behind this program is to teach the student how to operate the system rather than how the system operates. Progression through the course is governed by each individual's proficiency and rate of improvement during each session. End of course objectives were established for many knowledge and performance areas. Various modalities of training were selected to meet these objectives in the most realistic and efficient manner. As a result, many different testing methods are used to verify that training objectives have been met.

TRAINING PROGRAM

The training program was developed as the result of two studies accomplished by HumRRO. (Caro et al; Hall et al, 1969) As the simulator was being developed by Reflectone, Inc. to meet predetermined needs, the training program was being developed by HumRRO in cooperation with Coast Guard instructors already assigned to the aviation training division. Designing the training system concurrently with the simulator allowed modification of the simulator to accomodate changes in specific training goals. Both projects were consolidated when the Variable Cockpit Training System became operational in 1973.

One concept which provides a great deal of efficiency is the use of flight instructors as training managers. They conduct all phases of training and have the authority to modify course sequence to best suit the individual student. The training manager assures that the student meets the objectives for each phase of training prior to progressing to the next. As a result the student is continuously being tested against end of phase or end of course performance criteria.

TOTAL TESTING

In a program such as this where performance criteria are used for advancement through a course, a variety of evaluation methods must be used. Techniques used to measure phycho-motor performance would not be applicable

in the evaluation of mental processes. Responses requiring immediate recall must be tested in a different manner than those requiring research through official publications.

Oral Testing

Before going to the simulator the student must perform satisfactorily in the preflight briefing. This "briefing" consists of questions by the instructor regarding normal and emergency operation of the aircraft systems studied the previous night. If the student can verbalize his reactions to different malfunction indications, he is ready for the operational test in the simulator.

In addition, this dialectic method of briefing allows an evaluation of some aspects of troubleshooting strategy and other analytical thought processes that cannot easily be made in real time in the simulator. Each instructor works with two students who perform as a flight crew, responding to situations that have been previously covered in the briefing and study assignments. Students spend approximately half the time in the simulator as pilot and the other half as copilot. This relieves the instructor from his role as safety pilot and forces the student flight crew to handle their problems as they would in a real life situation.

Automated and Subjective Scoring

Prior to their first flight in the simulator a pair of students receive a mission briefing by the checkride instructor and plan an instrument flight to some predetermined airfield. This flight is objectively scored by means of a sub-program in the Datacraft 6024/3 computer which simultaneously operates both the HH-52A and the HH-3F simulators. Frequency and time out of tolerance are accumulated for each of twelve aircraft parameters and printed out as totals for each maneuver and for the entire checkride. An additional nine parameters pertaining to aircraft limitations are automatically displayed on the instructor's screen when the tolerable limits are exceeded.

During the checkride the instructor performs the duties of the air traffic controller, i.e., issuing clearances and providing weather information. In addition to normal procedures, malfunctions are automati-

cally inserted to allow the instructor to subjectively grade judgemental areas such as crew coordination, decision making and troubleshooting strategy. The instructor has the capability of inserting additional malfunctions if he deems it necessary.

Based upon overall performance, the instructor will make a subjective determination as to passing or failing the student. For the purpose of statistical analysis the instructor also issues a subjective grade based upon a four point scale. Objective and subjective scores were correlated (-1.0 = perfect correlation) with an obtained r of -.46 and -.56 for the two HH-3F checkrides and -.63 and -.70 for the two HH-52A checkrides. (Povenmire and Ballantyne, 1976) It is quite probable that better defined subjective criteria and grades for each task performed on the checkride would result in a higher correlation. It must be kept in mind, however, that subjective scores are based on an overall picture of the performance while objective computer scores are based on twelve selected parameters.

Debriefing

Debriefings are generally accomplished immediately after a procedure is performed. The instructor has the capability of freezing the simulator in any attitude or configuration simply by depressing a button on his panel. This enables him to point out and correct procedural or technical errors when they occur. In addition, the instructor has the option of playing back the previous five minutes of the flight in one minute intervals. During the playback the flight instruments and controls respond exactly as they did during the segment of the flight played back. Both features provide the capability of immediate feedback without the added distraction of manipulating the trainer.

During each simulator training flight, students are graded "satisfactory" or "practiced" on each procedure performed. Grades of outstanding or dangerous are also available for extreme cases and require an explanatory statement from the instructor. Areas graded as "practiced" are repeated until a satisfactory level of performance is attained. Post flight debriefings, as such, review problem areas from the simulator flight and preview the next day's lesson. When necessary, students are asked to review specific references or sound-on-slide presentations.

Written Exam

Due to the nature of their mission, Coast Guard helicopter pilots accrue very little actual instrument time in their day to day activities. Upon their arrival in Mobile, Proficiency Course students initially participate in a lecture-discussion on instrument procedures, rules, and regulations. This serves to refresh memories and to present information that has changed during the year. Knowledge of these and other pertinent facts will be tested either by performance in the simulator or on a written exam.

In January 1976, the decision was made to change the format of the annual instrument exam from the traditional multiple choice, right-wrong scoring method to a self scoring test which not only provided immediate information as to correct or incorrect response but feedback concerning why the response was right or wrong. The test was designed using a multiple choice erase-reveal scoring card with a coded reference for feedback information obtained from Van Valkenburgh, Nooger, and Neville. The purpose of the change was to utilize the annual instrument exam as a teaching tool and hopefully promote greater retention of the information presented. Studies have indicated that auto-instructional techniques are beneficial in terms of short range retention. (Jones, 1954; Lumsdaine and Glaser, 1960, pp. 52-93)

Our first attempt at creating a feedback type test consisted of the same type questions used on the previous exam but requiring a deeper level of knowledge. Based upon honest feedback from students and instructors it seemed obvious that the new test was extremely distasteful. Comments such as "nit picking;" "ambiguous," "irrelevant," "takes too much time," etc., were the prevelant opinions. Unfortunately, the comments were insufficiently lucid to provide the necessary insight for the problem.

At the suggestion of one of the instructors, key statements in the responses were changed to correct or incorrect instead of sorry about that, close but no cigar, wrongo, right on, etc. We also commenced making a statement during the Monday morning lecture-discussion describing the nature and purpose of the exam. Students were advised not to be concerned with the number of incorrect responses. They were told simply to research the incorrect answers as the purpose of the test is to teach not evaluate. In addition, correct answers requiring in depth analysis of reference mate-

rials were eliminated and negative type questions (select incorrect answer vs select correct answer) were removed where possible.

As a result of these changes student attitude toward the exam has changed completely. Comments such as "an excellent teaching aid besides being a test," "comprehensive, and I learned a great deal," "great improvement over past exams," "a good learning tool especially in the system employed," "it really makes you dig into the books," are indicative of the present feeling toward the exam.

We were naturally curious as to what effect this new attitude toward the test would have on the number of incorrect answers. With no real pressure it was anticipated that the mean number of errors would increase. A t test was performed on a random sample of 70 scores from each the original and the modified exams with no significant difference. (See Table 1)

	Original Exam	Modified Exam
Mean	76.97	75.38
SD	11.08	9.24

$t = .9193$ N.S.

TABLE 1

Comparision between mean scores for original and modified exams.

A modest correlation was found between error scores on the written exam and error scores on each of four automated checkrides ($r = .31$ to $.62$) indicating that those who score well on the written exam tend to do better in the aircraft. A study is presently under way to evaluate the long range retention value of this type of exam.

SUMMARY

When training objectives are defined in terms of behavioral change expected, the learning activity of the student can be better focused. Measurement of these objectives must be implemented in a manner appro-

priate to the expected behavior. With the change to self-scoring written exams, all facets of our evaluation process provide immediate feedback to the student. At any given moment he knows how he is progressing through the course.

BIBLIOGRAPHY

Caro, Paul W., Hall, Eugene R. & Brown, Gilbert E., Jr. Design and procurement bases for Coast Guard aircraft simulators. Ft Rucker, HumRRO Technical Report 69-103, Novmeber 1969.

Hall, Eugene R., Caro, Paul W., Jolley, Oran B. & Brown, Gilbert E., Jr. A study of U.S. Coast Guard aviator training requirements. Ft. Rucker, HumRRO Technical Report 69-102, November 1969.

Jones, R.S. Integration of instructional and self-scoring measuring devices. Abstr. doct. Dissert., Ohio State University, 1954, 65, 157-165.

Lumsdaine, A.A., & Glaser, R. *Teaching machines and programmed learning.* Washington: National Education Association, 1960.

Povenmire, H.K., & Ballantyne, K.M. Automated scoring of instrument flight checks. Orlando, *Proceedings, Ninth Naval Training Equipment Center/Industry Conference,* (in press), 1976.

Pressey, Sidney L. Teaching machine (and learning theory) crisis. *Journal of Applied Psychology,* Febuary 1963.

MEASUREMENT OF A PILOT'S CONTROL SKILLS IN AN INSTRUMENT FLIGHT MANEUVER

Brian D. Shipley, Jr.
US Army Research Institute

William V. Hagin and Vernon S. Gerlach
Arizona State University

INTRODUCTION

The Evaluation Problem

Student pilots must learn to operate a complex system in an unstable, frequently dangerous environment. These operational factors make stringent demands on an instructor pilot (IP) as he evaluates student pilot (SP) performance in an aircraft. The IP must monitor SP behaviors, unsafe performance conditions, and dangers in the airspace. Observations in all relevant areas of performance must be remembered or recorded and then used to arrive at a meaningful evaluation of performance skill.

Clearly, the IP must process large quantities of information. As training tasks become more complex or dangerous, demands on the IP tend to increase and they cannot adequately process all the essential information. As the information processing load increases, the integrity of safety procedures, training effectiveness, and evaluation methods may be compromised.

Based on the senior author's dissertation, the research covered in this report was completed at Arizona State University under the provisions of Air Force Office of Scientific Research Grant No. AFOSR 75-2900 to Vernon S. Gerlach.

Evaluation Methods in Pilot Training

After more than 30 years, rating scales remain the basic IP evaluation method because they best meet the needs of training management and the operational user - the IP - without intruding seriously into the IP's operational tasks (Koonce, 1974). As currently used, however, rating scale methods are marginally adequate for evaluations in training research and development because most observations using them lack sufficient discrimination; i.e., the observations tend to accumulate on two, or at best three points in the rating scale (Knoop & Welde, 1973). To make IP observations more discriminating in training research and development, extensive observer training and quality control programs must be developed and carried out (Horner, Radinsky, and Fitzpatrick, 1970; Koonce, 1974). This is a continuing and cumbersome process during the design and conduct of a study.

A variety of objective measures can be used with a flight simulator. The increasing sophistication of flight simulators allows greater opportunity to utilize objective data recording and processing methods. However, it is important that the objective measures be reliable, readily accessible, simply computed and, most of all, meaningful to users - IPs, student pilots, researchers, and others.

In this study two broad categories of objective measures are contrasted. An exact indicator was a measure easily observed at one or a few points in the sequence of events which comprise the profile of a training task. Most exact indicators used in training research are not well suited for general

applications because they are not usually derived from theoretical or general training requirements. Instead, they are selected or developed to reveal potential differences in performances which will answer some specialized research question. When that specialized need has been satisfied, all further interests in the measurements are dropped (Koonce, 1974).

By contrast, summary indicators, very popular in most training and research applications, reveal general trends or overall differences. A summary indicator is a global measure computed continuously or from a large number of discrete observations made throughout a single performance. Summary indicators have been criticized because they are usually insensitive to the effects of particular events or skills in a performance (Knoop and Welde, 1973) and because they are usually based on massive quantities of data with no definitely specified relationship to the evaluation problem (Christensen and Mills, 1967; Koonce, 1974).

The purpose of this study was to investigate exact indicators of particular skills or events that could be used to simplify objective measurement procedures in pilot performances; in short, to reduce excessive detail while not being uninformatively general (Youtz and Erickson, 1947). A step toward the solution of this detail versus generality dilemma would be achieved if a set of exact indicators, derived from a theoretical and/or training requirements analysis, were empirically shown to provide information to evaluate performances in a training research effort. Three separate tests of this hypothesis are reported. The research hypothesis was that a set of exact indicators would successfully replace a global summary indicator typically used in pilot training research.

APPROACH

Candidate Exact Indicators

Review of measures found in previous pilot training research produced three possibilities which might be effectively used as exact indicators. These three measures were the range, the maximum deviation from the standard, and performance time, all three of which were established in World War II as objective measures of flying skills (Miller, 1947). Performance time appeared particularly promising as a simplifying metric inasmuch as it had also been found to be a good indicator in two recent studies: Knoop and Welde (1973) and Brecke, Gerlach, and Shipley (1974).

The time and deviation measures actually used in this study were obtained from a previous research study on the Vertical S-A in which Brecke (1975) had automatically recorded measures of time, airspeed, heading, vertical rate, pitch, power, and altitude.[1] As a training task the Vertical S-A requires coordinated changes of pitch and power control. If each control change specified in the maneuver objectives is used to indicate a change of state, the standard Vertical S-A is composed of seven states (Figure 1).

Performance State Evaluation Models

Movements of an aircraft through space and time (i.e., the flight path) can be described with values from several variables such as altitude,

---

[1]The Vertical S-A is a complex instrument flight task performed in a flight simulator (USAF ATC Syllabus, 1975). Readers interested in more detail about the data should consult AFOSR Technical Reports #50430 and and #60229 by Brecke (1975) and Shipley (1976).

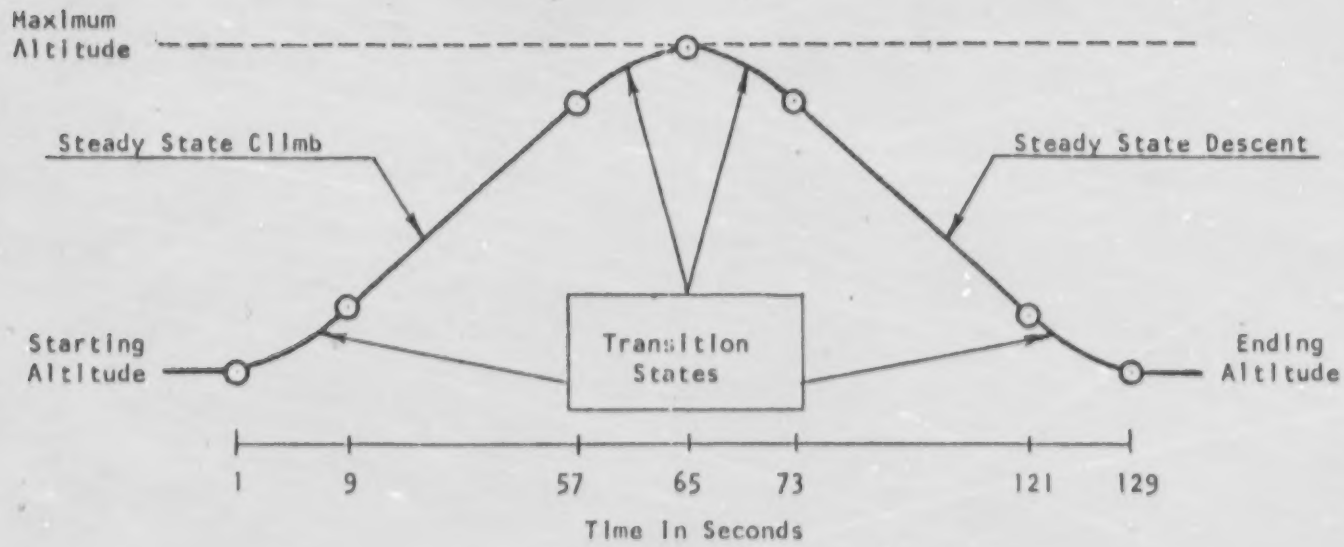


Figure 1.--Standard Vertical S-A Altitude Curve Showing Performance States and Performance Times

heading, airspeed, pitch, and power. To accomplish his mission, a pilot must control the movements of the aircraft by using information from some of these variables to make changes on others. Any set of values on these variables becomes a system "state". A state, then, is the complete set of variables needed to describe the entire performance of the pilot and the aircraft at any instant in time (Etkin, 1972). To evaluate a pilot's performance, values are observed in one or more states and the performance is assigned an appropriate score or mark to indicate the level of achievement.

Performance Time Algorithm

Since it is impossible for a performance to deviate significantly from standard times and still satisfy all other performance criteria throughout an entire flight, performance times can be used with deviations from flight path standards to simplify measurement and evaluation procedures for many maneuvers. In algorithmic models, if total time is not within given tolerances, some state or states must contain errors; these errors will be reflected as deviations from the standard. The state or states which contain the errors can be quickly located by examining the performance time for each state. For a state that fails the time test there are three possible outcomes: (a) if the entry conditions to that state deviate significantly from the standard, at least some source of the error will be in the preceding state or states; or, (b) if the entry conditions for a state are within the given tolerances, the sources of the error must be located in that state; and, (c) if the end point deviations are out of tolerance, the following state should also contain information about the pilot's performance.

INVESTIGATION I

The first investigation was carried out to compare "standard" time values for the Vertical S-A states with observed performances from two experienced pilots. The purpose was to discover whether or not the performance state evaluation model would reveal any significant differences between the performances of these two pilots. If the model detected significant differences between performances of experienced pilots, it should also make similar discriminations between performances of student pilots. Such discriminations are essential to effective evaluation in regular training and in experimental research on the effects of training methods.

Method

Two experienced pilots performed a sequence of 12 trials, 6 each, on the Vertical S-A in a flight simulator according to procedures described by Brecke (1975).

Procedures

Performance time and maximum altitude data were obtained from computer printouts of these performances and the means and variances were then computed for each pilot's data on total time, times for the seven performance states, and maximum altitude. Deviations of observed from corresponding standard values were also computed and used as the basis of tolerance limits in subsequent analyses of student performances.

Results

The means and variances of the performance times revealed three significant differences. There were two significant differences (.05) between means: a difference of 14.35 seconds on total time and a difference of 10.08 seconds on the time for the climb state. Finally, the variances were not homogenous for time in the third transition state.

INVESTIGATION II

The second investigation was carried out to determine whether total performance time and/or maximum altitude discriminated between performances of treatment groups in a training experiment. An *a priori* prediction, that differences among maximum altitude variances would discriminate among treatment group performances, was based on a training requirements analysis (Brecke and Gerlach, 1972; Gerlach, Brecke, Reiser and Shipley, 1972) and on the outcomes of a prior experiment (Brecke, Gerlach and Shipley, 1974).

Method

Thirty-nine student pilots were randomly assigned to one of five groups. Members of four experimental groups studied the objectives and different preflight instructions on how to perform the Vertical S-A; members of a control group studied only the objectives. After each subject had studied his assigned materials, he performed a sequence of six trials on the Vertical S-A according to procedures described by Brecke (1975).

Procedures

Total performance time values were analyzed with a two between- one within- subjects analysis of variance; repetition of performance trials on the Vertical S-A maneuver was the within subjects variable. Dunnett's method (Myers, 1966) was used to compare the means of the treatment groups with the means of the control group.

It was hypothesized that subjects in groups given experimental instructions would exhibit less variability among their performances at maximum altitude than subjects in either the control group or the groups given current instruction. To test this hypothesis, F-ratio tests for differences between group variances were used rather than tests for differences in means, i.e., ANOVA or t-tests (Winer, 1972).

Results

Total time. In the ANOVA on total time, there were significant main effects for type of instruction, number of response items in the instructional material, and performance trials. A significant interaction was also found between number of response items and performance trials. There were no significant differences on Dunnett's tests between any treatment group and the control group.

Maximum altitude. As predicted, the comparisons among the treatment groups' variances were significant on the maximum altitude measure. At maximum altitude, the variances of the two groups that received experimental instructions were significantly smaller (.05) than the variances

of the control group and the two groups receiving current instructions. The control group variance was not significantly different at maximum altitude than the variances of either group receiving current instruction.

Summary

In these first two investigations, performance times and maximum altitude discriminated among the performances of both individuals and groups. To the extent that the observed variations in performance time and maximum altitude reveal variability in the pilots' pitch and power control skills, these findings provide strong support for the hypothesis that exact indicators could be used as alternatives to summary indicators. In the third and final investigation, this hypothesis was examined by comparing the relationship between values on the two types of indicators.

INVESTIGATION III

In this last investigation, the hypothesis was that a set of exact indicator values would predict values of a summary indicator. If scores from a summary indicator in the Brecke (1975) data could be predicted with scores from a set of exact indicators, evidence would be obtained to support a recommendation to use the exact indicators in place of the summary indicator. A multiple regression analysis was used to test this hypothesis.

Method

Brecke (1975) used a summary indicator of variability in pilot performances called error amplitude. Error amplitude scores in Brecke's

study were computed with procedures described by Shipley, et al. (1974). Conceptually, error amplitude is a normalized root mean square measure obtained with the deviations of each observation from the corresponding standard and a set of tolerance limits for each parameter. In the present investigation, error amplitude scores were used as the criterion variable.

The exact indicators used as variates in the regression analysis were deviations from the standard on the performance state times, total time, and maximum altitude. Three combinations of performance state times were also included as variates in the regression design; these combinations were (a) the sum of the first three performance state times; (b) the sum of the last three performance state times; and (c) the sum of the times for all seven performance states. These sums were included in the design to test whether gross measures of time would be better predictors than more specific measures. Prior experience measures included as variates in the design were total hours as a pilot, flying time in undergraduate pilot training, total hours in the T-4 simulator, and number of minutes to complete the instructional materials.

Procedures

A set of 30 correlation matrices formed the basis of the regression analysis: one matrix for each of six performance trials across the five groups in Brecke's experimental design. The correlation coefficients were averaged over the five groups, after a Fisher's r to Z transform (Hays, 1963), and the means were then reconverted to equivalent r values. The result was a set of six matrices of average correlations, one matrix

for each trial. Finally, each of these matrices were submitted to a sequence of two stepwise regression analyses. The first analysis was based on the average correlations from the exact indicators and the sums; the second was based on the average correlations from the exact indicators and sums plus the prior experience measures.

Results

In the results of the regression analyses, support would be obtained for the hypothesis if two criteria were satisfied: (a) if a small set (4 to 6) of the 12 exact indicators was consistently selected in the regression equations; and (b) if equations using these consistently selected indicators would account for a substantial proportion of the variance (50% or more) in the criterion.

The regression analyses confirmed the hypothesis. It was found that 9 of 16 possible indicators and experience measures were frequently selected in the regression analyses. For each trial, various combinations from this set of nine were selected in regression equations that were limited to five predictor variables. These five variable equations accounted for at least 62% of the variance in the summary indicator, error amplitude.

The type of variable selected in these equations with five variables changed across the trials. For the first two trials, prior experience and combined times were predominant in the equation. By the last two trials, the exact indicators were predominant.

As measures of learning over trials, the summary indicator and the exact indicators were compared by making a plot of the means of each

variable across the trials. It was found that deviations from maximum altitude exhibited virtually the same learning trend over trials as error amplitude (r = .98).

These results provided convincing evidence to support the hypothesis that, in the Vertical S-A, exact indicators could be used to replace a summary indicator.

DISCUSSION

In these analyses, exact indicators were found to be more sensitive to the effects of different experimental treatments than was a summary indicator. In addition, when used alone, a set of exact indicators was found to account for moderate (34%) to large (82%) proportions of the variance in a summary indicator. These findings are provocative because they suggest that a few well chosen observations, i.e., fewer data points, can be used to obtain evaluations more economically than those obtained from typical summary indicators based on many data points.

The results of these investigations also support the use of a performance state evaluation model. Performance times and deviations from maximum altitude were found to discriminate reliably between both individual and group performances. Furthermore, these values enabled the researcher to identify the sources of errors within individual performances. It is interesting that performance time for the transition from climb to maximum altitude was consistently found among the predominant predictors in the regression analyses. In previous research, Gerlach, et al. (1972) predicted, on the basis of both theoretical as well as training requirements

analyses, that this location was the most likely source of student performance errors in the Vertical S-A. Coupled with the demonstrated sensitivity of deviations from maximum altitude, it is clear that the evaluation model could be effectively utilized for performance of the Vertical S-A in a flight simulator.

For the operational user, these findings suggest that exact indicators can be used to replace summary indicators and also global ratings. Maximum altitude, an example of an exact indicator in the Vertical S-A, was found to carry nearly all the information given by error amplitude, a summary indicator. Maximum altitude occurs at a well defined point in the Vertical S-A so that IPs can easily observe and use its value to replace a global rating. The dilemma of excessive detail versus uninformative generality in pilot training measurement and evaluation can be given a workable solution to the extent that similar values and their locations can be obtained for other pilot training maneuvers.

REFERENCES

Brecke, F. H. Cues and practice in flying training (Tech. Rep. No. 50430). Tempe: Arizona State University, College of Education, April, 1975.

Brecke, F. H., & Gerlach, V. S. Model and procedures for an objective maneuver analysis (Tech. Rep. No. 21201). Tempe: Arizona State University, Instructional Resources Laboratory, December, 1972.

Brecke, F. H., Gerlach, V. S., & Shipley, B. D. Effects of instructional cues on complex skill learning (Tech. Rep. No. 40829). Tempe: Arizona State University, College of Education, August, 1974.

Christensen, J. M., & Mills, R. G. What does the operator do in complex systems? Human Factors, 1967, 9, 329-340.

Etkin, B. Dynamics of atmospheric flight. New York: John Wiley, 1972.

Gerlach, V. S., Brecke, F. H., Reiser, R., & Shipley, B. D. The generation of cues based on a maneuver analysis (Tech. Rep. No. 21202). Tempe: Arizona State University, Instructional Resources Laboratory, December, 1972.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Horner, W. R., Radinsky, R. L., & Fitzpatrick, R. The development, test, and evaluation of three pilot performance reference scales (Tech. Rep. No. 70-22). Air Force Systems Command, Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, August, 1970.

Knoop, P. A., & Welde, W. L. Automated pilot performance assessment in the T-37: A feasibility study (Tech. Rep. No. 76-6). Air Force Systems Command, Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory, April, 1973.

Koonce, J. M. Effects of ground based simulator motion conditions upon prediction of aircraft pilot proficiency. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1974.

Miller, N. E. The problem of measuring flying proficiency. In N. E. Miller (Ed.), Psychological research on pilot training. Washington, D.C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

Myers, J. L. Fundamentals of experimental design. Boston: Allyn & Bacon, 1966.

Shipley, B. D. An automated measurement technique for evaluating pilot skill (Tech. Rep. No. 60229). Tempe: Arizona State University, College of Education, February, 1976.

Shipley, B. D., Gerlach, V. S., & Brecke, F. H. Measurement of flight performance in a flight simulator (Tech. Rep. No. 40830). Tempe: Arizona State University, College of Education, August, 1974.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

Youtz, R. P., & Ericksen, S. C. Analysis of the pilot's task. In N. E. Miller (Ed.), Psychological research on pilot training. Washington, D.C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

TELEVISION AS STIMULUS INPUT IN SYNTHETIC PERFORMANCE TESTING EXPERIMENT I

John Cockrell

Research Psychologist

21 July 1976

US Army Research Institute - Ft Knox Field Unit

Work Unit Leader	Donald M. Kristiansen
Field Unit Chief	Donald F. Haggard
Sr R&D Coordinator	MAJ Douglas Smith

TELEVISION AS STIMULUS INPUT IN SYNTHETIC PERFORMANCE TESTING: EXPERIMENT I

Over the past ten years or so the Army has made an effort to convert more and more of its testing to the "hands-on" performance mode. This is particularly true at training centers and at the beginning skill levels. Even more emphasis has been placed on performance testing in the last two or three years with the beginning of the Skill Qualification Testing (SQT) program. Performance testing is highly desirable because of its high face validity and high user acceptability; however, this type of testing is very costly, hard to standardize, and in many instances, not feasible.

The alternative to performance testing has generally been the standard, group-administered, knowledge-type, paper-and-pencil test. This type of test is easy to produce and administer but generally has low validity and low user acceptability.

About six years ago the Army Research Institute started a research program to investigate the possibilities of synthetic performance testing. Osborn (1970) did most of this early work and his efforts were devoted to finding a less costly type of performance test, but one that would still have reasonable validity and user acceptability.

Osborn coined the phrase "synthetic performance test" to refer to any performance test that was less than a full "hands-on" test, but more than the group-administered, knowledge-type, paper-and-pencil test. Included under synthetic performance tests were all tests that used any type of simulated inputs or responses. Also included under synthetic performance tests were part-task tests in which only one or a

few response components of a task were measured.

Put Slide 1 about here

Osborn's conceptual scheme is based on two general ideas. The first of these is that as one goes up the scale of stimulus and response fidelity, it is possible to achieve more and more validity, but at the same time the costs get greater and greater and therefore feasibility goes down. The problem here is to try to find a trade-off validity-feasibility point that would be useful and acceptable.

Put Slide 2 about here

The second idea is that each task can be broken down into cognitive, perceptual, and motor components, and that performance on each of these components can be tested in isolation. What is more important is that many of these components can be tested by synthetic testing methods. If it can be demonstrated that performance on one or a few response components correlates highly with performance on the entire task, then we have a much less costly method of measuring performance on this task.

Put Slide 3 about here

As an offshoot from Osborn's work, the Army Research Institute started a research program about two years ago to investigate the use of audio-visual media in synthetic performance testing. The reasoning behind this research program is that audio-visual media stand somewhere in the

medium stimulus fidelity range, and at the same time, are at the medium to high end of the feasibility scale. Thus, it is possible that audio/visual media may represent a good validity-feasibility tradeoff point insofar as stimulus fidelity is concerned.

Several experiments in this research program are now in process. All of the different types of audio/visual media are being investigated. The first of these experiments was concerned with television as the stimulus input and this paper presents the results of this experiment.

OBJECTIVES

The primary objective of the experiment was to explore the possibility of using television as the stimulus input in performance testing.

Put Slide 4 about here

Specifically, answers were sought for the following:

1. Feasibility of television testing
   a. Understandability of the problems
   b. Ease of responding
   c. Time allottment for responding
   d. Production and administration costs
2. Acceptability of television testing by examinees
3. Comparability of television test to "hands-on" test and paper-and-pencil test.

METHOD

The overall method consisted of (1) producing a television test for the job field of tank crewman (11E MOS), (2) producing a parallel paper-and-pencil test covering the same items, and (3) comparing the results made on these two tests with the results made on an existing "hands-on" performance test which covered many of the same items.

The job field of tank crewman was selected because much prior research had been done in this field. A complete task analysis was available and a "hands-on" performance test has been in use for the Tank Crewman Advanced Individual Training course for two years. It was felt that this existing "hands-on" performance test would be a good base against which to compare the television and paper-and-pencil tests.

Television Test

The first step in producing the television test was to select approximately 50 performance measures from a number of critical tasks for the Tank Crewman MOS. These performance measures had been developed previously in a study by Ford, et al (1974).

From these performance measures a shooting script was produced and the problems were televised. The shooting and editing required approximately 30 calendar days of which about 15 were actual working days.

The test included 48 problems plus four practice problems and had a running time of 53 minutes. The problems were wide ranging over many tasks in the particular career field.

The final test can be considered a "performance" test in that all of the problems were adapted from a "hands-on" performance test. No pure knowledge-type items were included and each problem was directly relevant to the normal tasks which the examinees are called on to carry out in their work. Each problem covered only one response component in a particular task such that it was necessary to include a series of problems in order to adequately cover some of the more important tasks. The problems were presented in roughly the same sequence as the tasks would be performed on the job; however, with only 48 problems, the test only covered a small fraction of the total job.

Each problem was approximately 60 seconds in duration. The first 50 seconds was used to set the scene and pose the question. The last ten seconds was the time allotted for the response. The test ran continuously and each problem had to be answered in the allotted ten seconds. There was no opportunity to study any of the problems and there were no second chances in responding.

Responding Device

One of the reasons for producing the television test was to check out a new electronic response device which could be used with television. This device permitted the examinees to respond to the problems by touching the images on the screen with a stylus or a plastic gun-sight reticle. The correct answer to each problem was programmed on the test tape and the device recorded "right" or "wrong" according to where and when the examinee touched the screen.

As might be expected this electronic response device did not work. However, it was possible to proceed with the experiment and test out the concept of the device by using human graders. These graders sat in back of each examinee and recorded "right" or "wrong" according to where and when the examinee touched the screen.

Problem Types

Three problem types were used on the television test with some variation within each type. These were:

Put Slide 5 about here

---

1. MULTIPLE CHOICE. The examinees responded by touching one of the choices shown on the television screen. The choices were either words or pictures of equipment.

2. ERROR DETECTION. The examinees responded by touching the television screen at time and place of error .

3. MOTOR MANIPULATION. The examinees responded by manipulating a gun-sight reticle on the screen, simulating aiming main gun at stationary or moving targets.

Comparision Tests

Two other tests were used in addition to the television test for comparison purposes. The first of these was a paper-and-pencil test and the second was a "hands-on" performance test.

1. Paper-and-Pencil Test. The paper-and-pencil test paralleled the television test on a problem by problem basis. This test was composed of pictures, diagrams, drawings, and printed words. All questions were multiple choice and a separate answer sheet was used.

2. "Hands-on" Performance Test. The "hands-on" test was one that was routinely administered to tank crewmen trainees as a final examination for the Advanced Individual Training course. This test was given in the form of a county fair with 8 stations and 30 performance measures. The examinees were graded on a "go-no-go" basis for each performance measure. For each "no-go" the examinees were required to seek out remedial training and report back for a retest. If the retest was a no-go the examinee had to report back the next day after further remedial training for one final test. For the purposes of the present experiment, the score recorded for each examinee was the number of first-round no-go's. It was recognized that this was not a particularly good criterion measure, because the number of no-go's was very small. However, as this first television experiment was considered an exploratory experiment, it was decided to make the comparison anyway.

Examinees

The examinees have just completed the "hands-on" performance test and had a three-day wait for graduation and processing to the next assignment. During this three-day wait they were available for participation in the experiment. A total of 134 examinees was used.

Procedure

The television test was administered to four examinees at a time using four television monitors and one tape player. The monitors were arranged such that no examinee could see the responses of the others. Each examinee was assigned a grader whose job was to note if the examinee touched the correct spot on the screen. The grading job was quite simple and consisted of marking an answer sheet "right" or "wrong" for each

problem. Prior to the start of the experiment the graders had been trained.

A pilot run of the experiment was conducted with eight examinees in order to practice the procedure and grading. The procedure was smooth and no difficulties were encountered.

Each group of 16 examinees reported for the morning session or the afternoon session and were randomly divided into a television group and a paper-and-pencil group. All 8 of the paper-and-pencil group were administered the paper-and pencil test as soon as the orientation was completed. The television test was administered to four examinees at a time and the remainder of the group was assigned to a small room to wait for the next round of testing. Both the television test and the paper-and-pencil test required approximately one hour to complete.

Approximately ten minutes of training was required to teach the examinees to respond to the television problems. Using the stylus was no problem, but there was some difficulty in using the plastic reticle. The examinees were trained by having them respond to the four practice problems. If the responses were not perfect, the tape was stopped and the four practice problems were run again. In no case was it necessary to run the practice problems more than twice.

RESULTS

Feasibility of Using Television in Testing

The examinees did not appear to have any difficulty in understanding the problems, and the responding proceeded smoothly for most of the problems. The examinees responded very quickly on the easy problems and on difficult problems, the amount of response time allotted

(10 seconds) did not appear to be a problem, although there would usually be a lot of hovering over the answers. Only on a few problems did the examinees fail to respond. When queried after the completion of the test about the amount of response time, most of the examinees indicated that for the most part the response time was adequate. A few examinees said that more response time should have been allotted to some of the problems.

The response method did not appear to pose a problem. Responding with the stylus was very natural, and responding with the gun reticles quite familiar to the examinees. During the preliminary narration for each problem, the narrator would instruct the examinees on the proper reticle to use. Occasionally, one of the examinees would have some difficulty finding one of the reticles.

The administration of the television test was somewhat cumbersome. Each group of four examinees had to be instructed for about ten minutes on the response method and it was necessary to not only give the instructions but also to demonstrate the use of the reticles and to permit the examinees to practice the responses. Instructions and practice were also necessary on adjustments that could be made to the television sets. The administration could be made more feasible by increasing the number of television monitors but it would be still be advisable to have one test administrator for each four examinees due to the unfamiliarity of the response method.

Acceptance of Television Testing

The reaction of the examinees to the television test appeared to be quite favorable. Most of the examinees actually preferred the television test to the "hands-on" test and all of the examinees thought the television test was fair. Even when queried about the test being used as a basis for promotion or extra pay they still thought the test was fair. Some of the examinees preferred the "hands-on" mode of testing but no one preferred the paper-and-pencil mode. Some of the reasons for preferring the television test were as follows:

1. Scoring is fairer and not dependent upon the whims of the test administrator.

2. Testing is faster and not so drawn out.

3. In television testing no one is shouting at you and ordering you around.

The examinees also indicated that television testing would be better than paper-and-pencil testing because the questions would be more understandable and would require much less reading.

Comparison of Television, Paper-and-Pencil, and "Hands-on" Tests

The comparison between the mean percent error made on the television test and that made on the paper-and-pencil test is shown in the next slide. The means for the television test and paper-and-pencil test

---

Put Slide 6 about here

---

do not differ to any great degree indicating that the difficulty of the two tests are fairly equal.

One interesting facet of the data indicates that the television test appears to be much more difficult in the afternoon than in the morning. There does not appear to be any morning-afternoon effect for the paper-and-pencil test.

The analysis of variance shown in the next slide shows no difference between the television test and the paper-and-pencil test in terms of diffculty. There was a significant morning-afternoon effect but the more meaningful result is the significant MS interaction. Analysis of this MS interaction reveals that the morning-afternoon effect is concentrated on the television test and not on the paper-and-pencil test.

Put Slide 7 about here

In order to check on whether the afternoon examinees may have been less qualified that the morning examinees, the first round no-go's from the "hands-on" test were analyzed. These results are shown in the next slide. Inspection of the means indicates little difference

Put Slide 8 about here

between the television and paper-and-pencil groups, and also little difference between the morning and afternoon groups. If anything the afternoon group performed slightly better than the morning group. An analysis of variance of these results showed no significant difference for any of the variables.

Although there was no difference between the overall scores on the television and paper-and-pencil tests, it is possible that there might be differences among the various items. Accordingly, the items

were grouped by response type (multiple-choice, error detection, and motor manipulation) and Chi-square tests were computed for each item. The next three slides show these results. For the multiple choice

---

Put Slides 9, 10, and 11 about here

---

items there was little differnce between the television and paper-and-pencil versions. Only one item out of 13 showed a significant difference. For the error detection items there was a substantial difference with six out of nine items showing a significant difference. It is interesting to note that errors of commission are more difficult to detect on television whereas errors of omission and no-error items are more difficult to detect on paper-and-pencil. Three of the eight motor manipulation items show some significant difference and all three of these items show more difficulty for the television test. The net result of this item difficulty analysis shows five items that are more difficult on television and five items that are more difficult on paper-and-pencil. This cancelling effect is reflected in the overall non-significant difference between the television test and the paper-and-pencil test.

The last analysis shows the correlation between the television test and the "hands-on" test, and between the paper-and-pencil test and the "hands-on" test. These correlations are also broken down for the morning and afternoon groups. The next slide shows these results.

Put Slide 12 about here

There is a low positive correlation between the television test and "hands-on" test. The paper-and-pencil correlation is significantly different from zero; however, there is no significant difference between the television test and the paper-and-pencil test. The breakdown for morning and afternoon groups show a somewhat higher positive correlation for the afternoon group and very little correlation for the morning group. Once again, there is no significant difference between the television and paper-and-pencil tests.

DISCUSSION

The results from this exploratory study indicate that it is possible to produce a synthetic test using television as the stimulus input. The examinees can understand the problems, make the proper responses, and the time limit of ten seconds for the response is generous for those examinees who know the correct answer.

The experience gleaned from the production and administration of this prototype test indicate that television testing is somewhat cumbersome and costly. However, in all fairness to television, it should be mentioned that television testing is far less cumbersome and costly than "hands-on" testing. The production of the tape from conception to final editing required several months and utilized the services of a substantial number of professional people. Television tests are also somewhat inflexible, not only in terms of the difficulty in effecting changes in the test, but also in terms of the timing

decisions that have to be made prior to the production of the test. The timing decisions have to do with the amount of time to allot to posing the question and the amount of time to allot to the response. Television testing will have a much more promising future if a presentation and response device can be designed which will permit the examinee to advance to the next problem as soon as the present one is answered, to see the same problem twice, to change answers to a problem, and to review the entire test. Such a capability would permit the flexibility of presenting multi-part problems such as in troubleshooting, and the presentation of multi-media problems such as using both television and technical manuals in the same problem.

There is evidence from the present experiment that television testing is highly acceptable to the examinees. The predominant attitude of the examinees was that the test was little different from the "hands-on" tests in the Advanced Individual Training course except that television was quicker and less subject to error in the scoring. All of the scenes were quite familiar to the men and the problems were ones that they had been studying for 8-13 weeks. One gets the impression from talking to these tankers that they consider their job to be primarily a cognitive one rather than one involving motor skills.

When television is used in the multiple-choice format there is little evidence to show that it offers any advantage over a slide format or a paper-and-pencil format. Prior to the start of the experiment it was felt that television would offer an advantage for

those problems where motion was an integral part of the stimulus. However, this proved to be true for only one multiple-choice item and was not true for two other items. Perhaps, if more motion-type items had been included in the multiple-choice category, the advantage might have been shown.

In the error-detection category there did appear to be a clear-cut difference between television items and paper-and-pencil items. Here the fidelity of the stimulus did seem to play a role and the enriched stimulus of the television picture may have presented cues to the examinees. The two error detection items that proved to be more difficult for television were two of the first error detection items to be presented to the examinees. Since error detection was an unfamiliar response to the examinees, this unfamiliarity may have caused some difficulty. This same phenomena can be seen in the motor manipulation items which is an even more unfamiliar response. Here the examinees had more difficulty with the first few items on the television test than on the paper-and-pencil test.

The correlations between the synthetic tests and the "hands-on" test are too low to warrant recommending the substitution of these tests for the "hands-on" test. However, the correlations for the afternoon groups are high enough to encourage further research. The criterion test used in the present experiment was somewhat unsatisfactory and future research will attempt to develop a more suitable criterion test.

The drop in the scores on the television test for the afternoon group as compared to the morning group was interesting but unexpected. One possible explanation for the drop may lie in the fact that the examinees were required to stare continuously for approximately one hour at a fairly large television screen from a very close distance. The human being may be able to tolerate this type of strain in the morning but the accumulated fatigue of the afternoon plus the usual heavy Army lunch may have combined with the strain to produce a let down.

CONCLUSIONS AND RECOMMENDATIONS

The overall conclusion from this experiment is that television testing is feasible. The tests can be produced at a reasonable cost, can be administered in a reasonable manner, and are understandable by the examinees. Television tests also have high acceptability with the examinees.

The evidence with regard to validity is inconclusive. The present experiment did not have a criterion test that was satisfactory and therefore no firm conclusions can be drawn with regard to validity. Some of the correlations were high enough to encourage further research.

The evidence did not demonstrate the superiority of television over paper-and-pencil testing. The overall error scores were comparable. However, the item analysis did reveal some major differences between television and paper-and-pencil. Further research is needed to determine if these differences are reliable and are related to the nature of the medium.

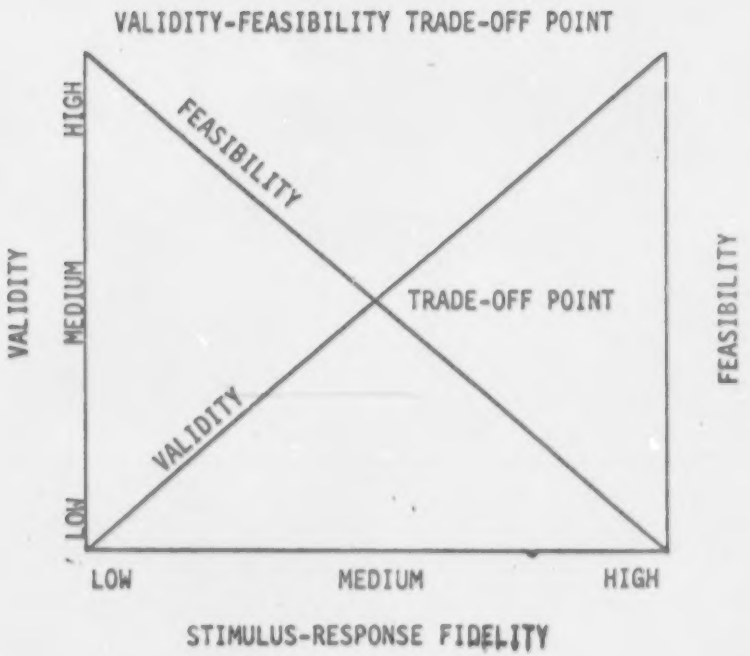
It is recommended that research on television testing be continued with emphasis on producing a precise test and validating it against a precise criterion.

REFERENCES

Ford, J. Patrick, Harris, James H., and Rondiac, Peter F. Performance measures for AIT Armor Crewmen. HumRRO Consulting Report, April 1974.

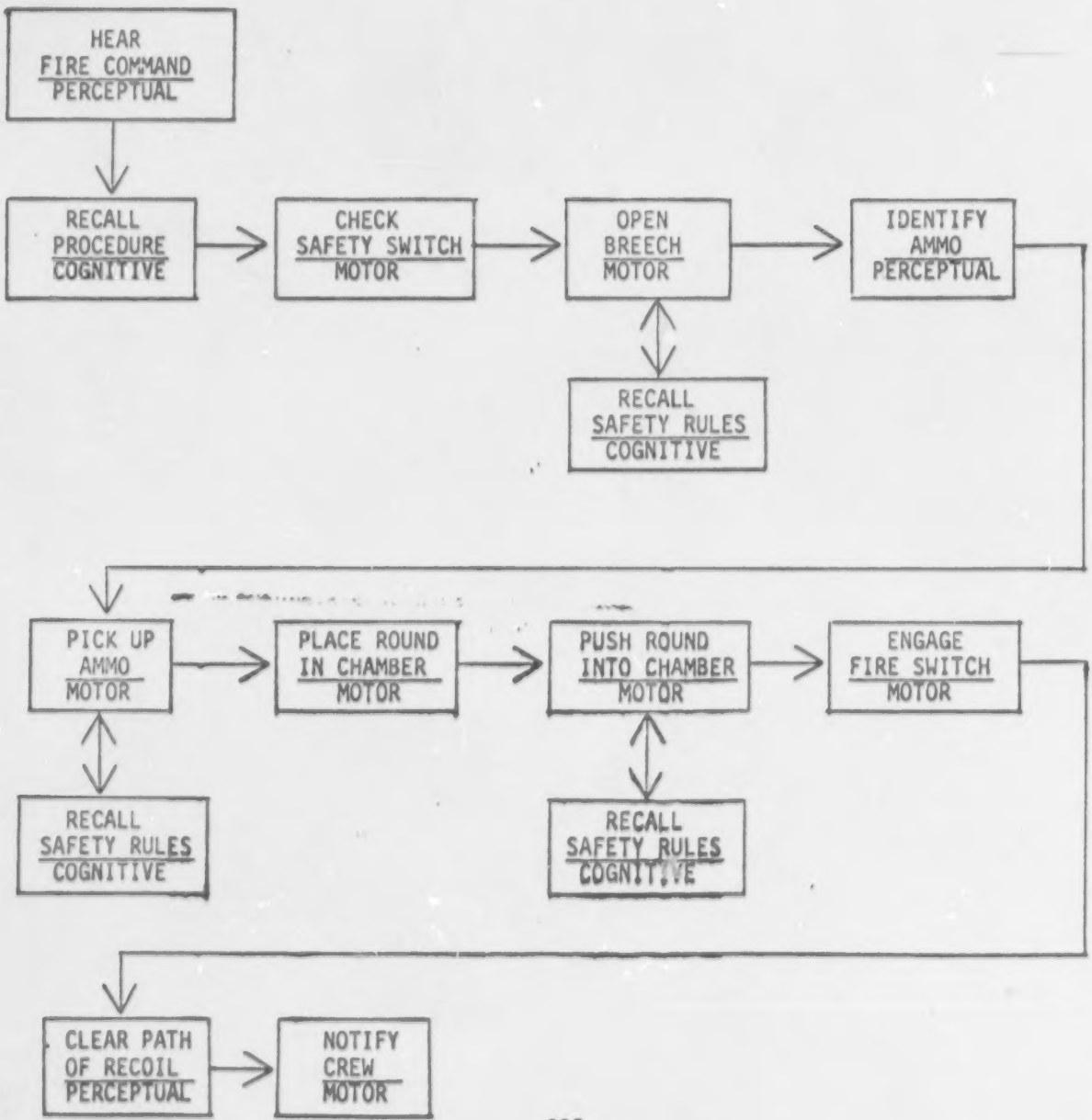
Osborn, William C. An approach to the development of synthetic performance tests for use in training evaluation. Paper presented at 12th Annual Military Testing Association Conference, French Lick, Indiana, HumRRO Professional Paper 30-70, December 1970.

SLIDE 1


VALIDITY-FEASIBILITY TRADE-OFF POINT
VALIDITY
HIGH
MEDIUM
LOW
FEASIBILITY
VALIDITY
TRADE-OFF POINT
FEASIBILITY
LOW
MEDIUM
HIGH
STIMULUS-RESPONSE FIDELITY

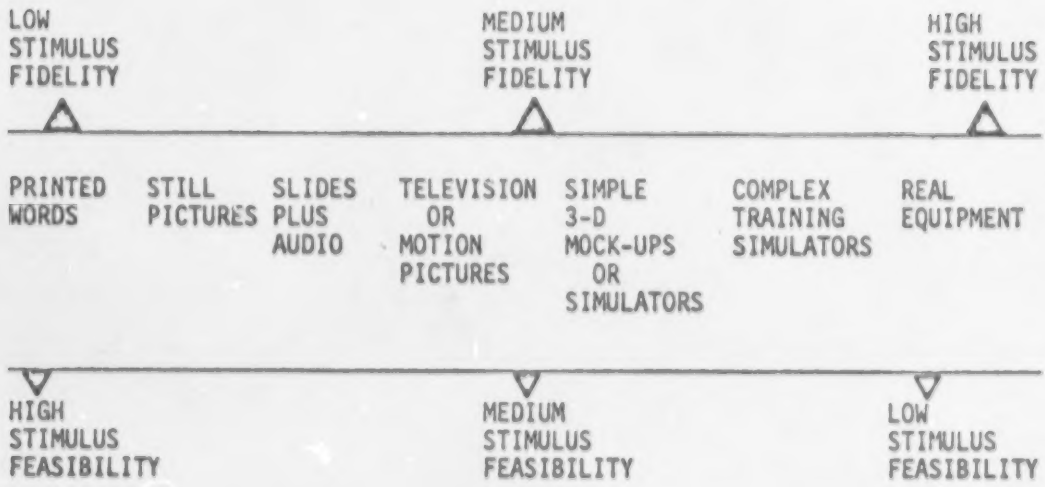
SLIDE 2

RESPONSE COMPONENTS FOR TASK: LOADING ROUND INTO MAIN GUN



SLIDE 3

STIMULUS FIDELITY AND FEASIBILITY



SLIDE 4

OBJECTIVES

1. FEASIBILITY OF TELEVISION TESTING
2. ACCEPTABILITY OF TELEVISION TESTING BY EXAMINEES
3. COMPARABILITY OF TELEVISION TEST TO "HANDS-ON" PERFORMANCE TEST AND PAPER-AND-PENCIL TEST

SLIDE 5

PROBLEM TYPES ON TELEVISION TEST

1. MULTIPLE CHOICE

2. ERROR DETECTION

3. MOTOR MANIPULATION

SLIDE 6

MEAN PERCENT ERRORS MADE ON TELEVISION AND PAPER-AND-PENCIL TESTS IN THE MORNING AND AFTERNOON TESTING SESSIONS

	TV	P-P	MEAN
MORNING GP	16.46	24.23	20.34
AFTERNOON GP	27.54	26.05	26.80
MEAN	23.74	25.42	

SLIDE 7

ANALYSIS OF VARIANCE FOR TELEVISION
AND PAPER-AND-PENCIL TESTS

SOURCE	df	MS	F	p
TV VS. P&P (METHOD)	1	9.50	1.04	ns
MORNING VS. AFTERNOON (SESSION)	1	124.57	13.60	<.01
MS	1	58.66	6.40	<.02
WITHIN CELL	130	9.16		

SLIDE 8

MEAN PERCENT FIRST-ROUND NO-GO'S MADE ON
HANDS-ON PERFORMANCE TEST

	TV	P-P	MEAN
MORNING GP	8.16	9.08	8.62
AFTERNOON GP	6.52	8.48	7.50
MEAN	7.16	8.68	

SLIDE 9

COMPARISON BETWEEN THE INDIVIDUAL MULTIPLE-CHOICE ITEMS ON THE TELEVISION AND PAPER-AND-PENCIL TESTS (PERCENT ERROR)

ITEM#	TYPE ITEM	TV TEST	P-P TEST	X^2	p
1	MULTIPLE CHOICE	1	0	-	ns
2	MULTIPLE CHOICE	0	2	-	ns
4	MULTIPLE CHOICE	16	12	-	ns
5	MULTIPLE CHOICE	17	16	-	ns
6	MULTIPLE CHOICE	43	33	1.43	ns
7	MULTIPLE CHOICE	21	14	.82	ns
8	MULTIPLE CHOICE	38	72	13.83	<.01
9	MULTIPLE CHOICE	47	52	-	ns
10	MULTIPLE CHOICE	26	20	-	ns
11	MULTIPLE CHOICE	3	5	-	ns
12	MULTIPLE CHOICE	0	0	-	ns
17	MULTIPLE CHOICE	1	3	-	ns
29	MULTIPLE CHOICE	81	77	-	ns

SLIDE 10

COMPARISON BETWEEN THE INDIVIDUAL ERROR-DETECTION ITEMS ON THE TELEVISION AND PAPER-AND-PENCIL TESTS (PERCENT ERROR)

ITEM NUMBER	TYPE ITEM	TV TEST	P-P TEST	X^2	p
3	ERROR DETECTION (COMMISSION)	51	28	7.47	< .01
11	ERROR DETECTION (COMMISSION)	11	11	-	ns
14	ERROR DETECTION (COMMISSION)	40	12	11.09	< .01
30	ERROR DETECTION (COMMISSION)	46	56	1.42	ns
16	ERROR DETECTION (OMISSION)	19	52	14.88	< .01
18	ERROR DETECTION (OMISSION)	4	47	18.90	< .01
26	ERROR DETECTION (OMISSION)	13	45	11.26	< .01
15	ERROR DETECTION (NO ERROR)	16	19	-	ns
20	ERROR DETECTION (NO ERROR)	17	66	30.36	< .01

SLIDE 11

COMPARISON BETWEEN THE INDIVIDUAL MOTOR MANIPULATION ITEMS ON THE TELEVISION AND PAPER-AND-PENCIL TESTS (PERCENT ERROR)

ITEM#	TYPE ITEM	TV TEST	P-P TEST	X^2	p
19	MOTOR MANIPULATION	46	16	12.24	<.01
21	MOTOR MANIPULATION	51	17	16.12	<.01
23	MOTOR MANIPULATION	27	11	3.75	<.10
24	MOTOR MANIPULATION	11	2	-	ns
25	MOTOR MANIPULATION	14	11	-	ns
27	MOTOR MANIPULATION	23	34	-	ns
28	MOTOR MANIPULATION	23	28	-	ns

SLIDE 12

CORRELATIONS BETWEEN THE SYNTHETIC TEST SCORES (TELEVISION AND PAPER-AND-PENCIL) AND THE HANDS-ON PERFORMANCE SCORES

OVERALL TEST	r	p	
TELEVISION VS HANDS-ON	.24	ns	DIFFERENCE IS NON-SIGNIFICANT
PAPER-AND-PENCIL VS HANDS-ON	.33	<1	
MORNING GROUP			
TELEVISION VS HANDS-ON	-.09	ns	DIFFERENCE IS NON-SIGNIFICANT
PAPER-AND-PENCIL VS HANDS-ON	.16	ns	
AFTERNOON GROUP			
TELEVISION VS HANDS-ON	.47	<.01	DIFFERENCE IS NON-SIGNIFICANT
PAPER-AND-PENCIL VS HANDS-ON	.40	<.01	

Sequential Analysis for Individual Proficiency Decisions[1]

Claramae S. Knerr and Kenneth I. Epstein

US Army Research Institute for the Behavioral and Social Sciences

The measurement of individual proficiency is required in many types of situations. One example occurs during individualized instruction, or other training where criterion-referenced tests are administered, and where the scores are interpreted to indicate whether or not the examinee has mastered the skill being trained and tested. The interpretation of interest is the classification of examinees into proficiency categories. For example, examinees classified as failures may be assigned remedial training. Typically, the test contains a given number of items and a minimum passing score, or cut-off score, is stated as a percentage of the items that the examinee must answer correctly in order to be classified as proficient. Frequently, 80% is set as the minimum passing score, while 100% correct may be required for critical skills. When the number of test items is small, there is known to be a large amount of error in the classification decisions made on the basis of the test score (Novick & Lewis, 1974). Error refers to one of two situations: first, the decision that examinees are not proficient when in fact they have mastered the skill; and second, the decision that examinees are proficient when in fact they have not mastered the skill.

The Army has been using a procedure for training evaluation, called sequential analysis, which offers an effective solution for the classification of individuals into proficiency categories. The application of sequential analysis to the evaluation of instructional materials has been described elsewhere (Epstein, 1975; Jeter & Singleton, 1975, US Army Training and Doctrine Command, 1975). The sequential analysis procedure has also been applied to the evaluation of individual proficiency levels (Kriewall, 1968; Linn, Rock, & Cleary, 1972).

Sequential analysis is an alternative to traditional test situations where the number of test items is fixed before the test is administered, and no decision is made until all of the items have been analyzed. The use of a sequential testing strategy takes advantage of the notion that either very high or very low proficiency can be detected with a small number of items, and that more extensive testing is necessary only for individuals of borderline proficiency. Sequential analysis calls for observing one response at a

---

[1] The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

time with the possibility of a proficiency classification decision after each observation. That is, a test item is administered, and based on the examinee's response, proficiency is either accepted (pass for the whole test), rejected, or no decision is made. For the case where the proficiency decision cannot be made, another item is administered and the decision rule is applied again. The process continues until the proficiency classification can be determined.

Sequential analysis requires that the examinee respond to only as many items as are necessary to insure that the expected error in the classification decision does not exceed some specified value. Prior research (Linn, Rock, & Cleary, 1972) has shown that sequential analysis requires approximately half as many items as conventional tests (of fixed length) to achieve the same level of accuracy. The strength of the sequential analysis model, therefore, lies in efficient test length combined with control over the extent of decision error.

To use the sequential model, four parameters are set: a high cut-off score for definite proficiency, a low cut-off score for definite failure, and two allowable error rates (one for each cut-off score). These parameters are used to calculate the lines to be drawn on a graph, as shown in Figure 1. Scores are plotted for one examinee at a time. The examinee's first item score is plotted one unit to the right of the zero point if the response is correct. Figure 1 shows the graph for an examinee who performed correctly on the first ten items. The line representing the scores lies directly across the bottom of the graph, to the right from the zero point, and it crosses into the region labelled "pass". This examinee was classified as proficient at the task, given the first ten items. Figure 2 shows the graph for an examinee who failed the first five items. Each incorrect response is plotted one unit to the right and one unit up from the zero point. The line rises until it crosses into the region labelled "fail". This examinee was classified as not proficient at the task, given the first five items. Figure 3 shows the more usual case, where the examinee passed some items and failed others. The examinee performed incorrectly on item 5, and had to be scored on fifteen items before the proficiency decision was made.

These figures show that the examinee can fail in the first five items, but must respond correctly on at least ten items to pass. These different pass and fail requirements are produced by the four parameters that determine the bounds of the pass and fail regions. Two of the parameters are cut-off scores. For example, one way to restate the traditional 80% correct criterion in sequential testing terms is that examinees who perform correctly 90% of the time are considered proficient, while examinees who perform correctly only 70% of the time are considered not proficient. The proficiency level of examinees who perform between 70% and 90% defines a zone of indifference. That is, for practical purposes, either a mastery or a nonmastery classification is acceptable.

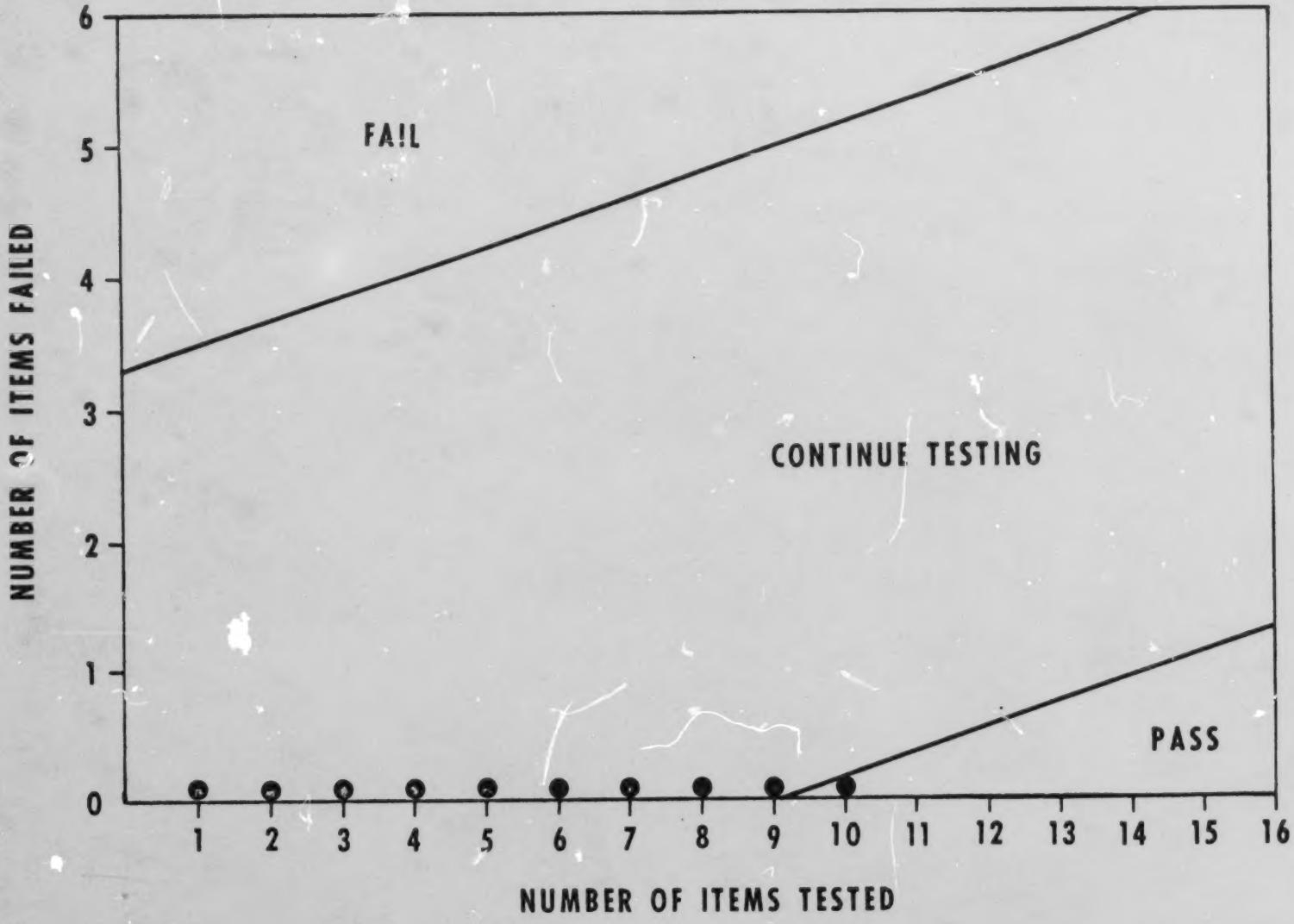
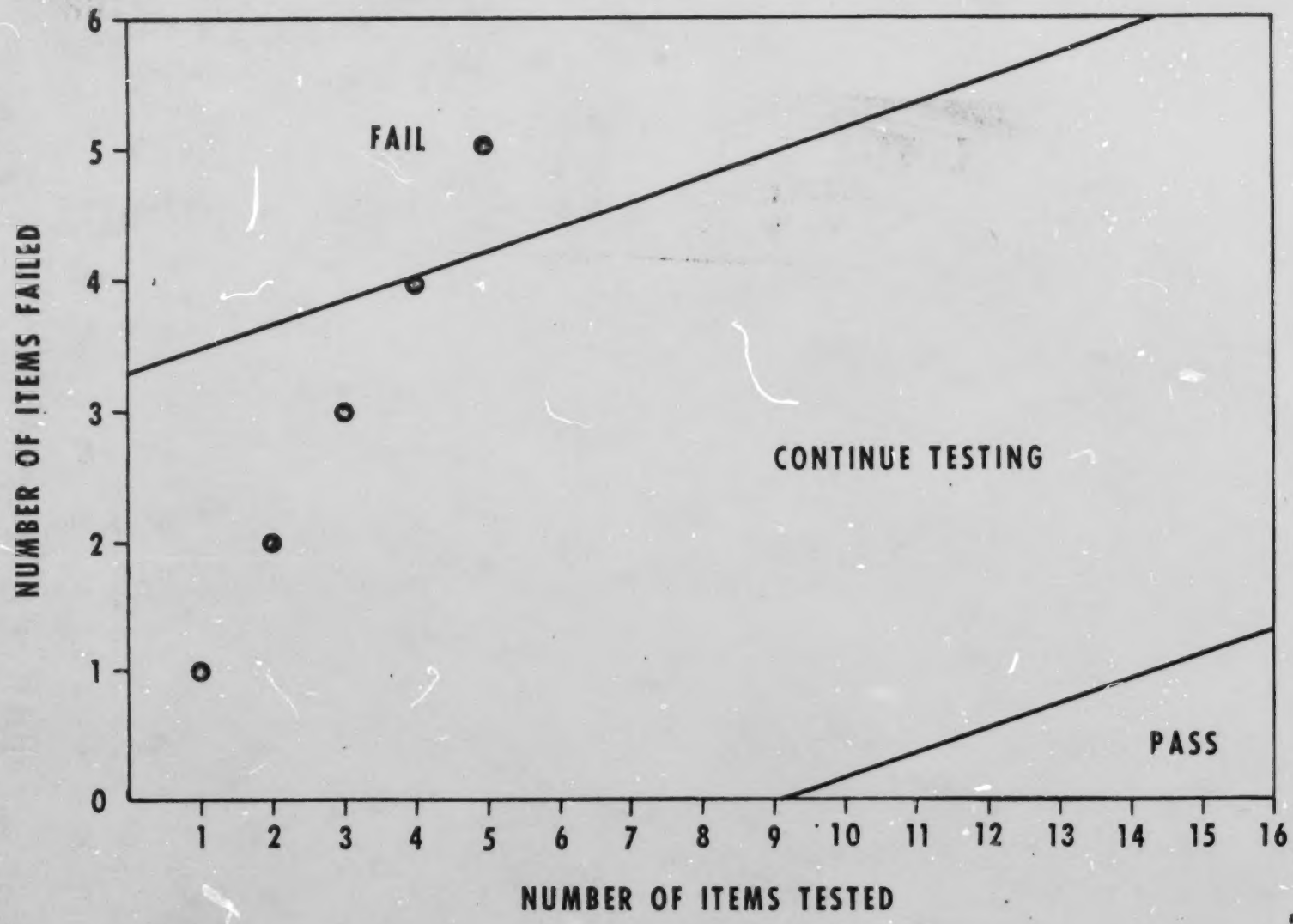


FIGURE 1. MASTERY DECISION GIVEN 10 ITEMS



293

294

FIGURE 2. NONMASTERY DECISION GIVEN 5 ITEMS

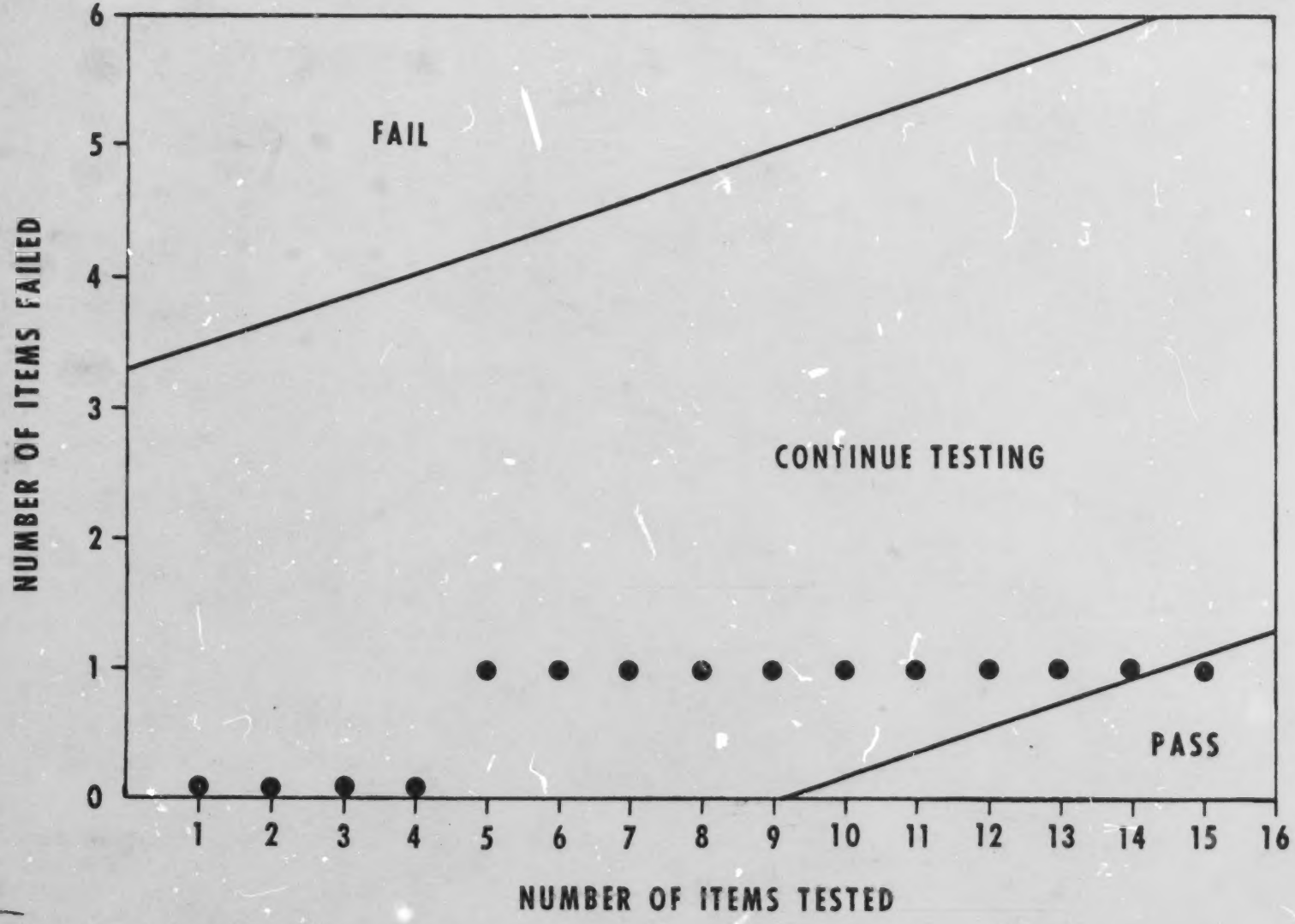


FIGURE 3. MASTERY DECISION GIVEN 15 ITEMS

The high and low cut-off criteria (90% and 70%) refer to population values, which are not known with absolute accuracy from the sample data when the classification decision is made. Two other parameters, alpha (α) and beta (β) are used to specify the extent of allowable classification error. The value of alpha is determined by answering the question: What percentage of truly proficient examinees can be erroneously failed? In the present case, erroneously failing examinees who were actually able to perform the task was considered to be very costly (e.g., erroneously failing skilled soldiers would discourage them, and would lead to expensive, unnecessary remedial training). Thus the value of alpha was set at .01 to keep this type of error very low. The value of beta is set by answering a similar question: What percent of incapable examinees can be erroneously passed? In the present case, the soldiers had access to further training on the tasks after the research, and since they could correct deficiencies later, the cost of incorrectly passing them was thought to be low. Thus, beta was set at .10. The number of items required for the decision depends on the alpha and beta levels and the cut-off scores. No more items will be administered than are necessary to keep classification errors within these pre-determined bounds.

The mathematics and computational procedures are available elsewhere (Crow, Davis, & Maxfield, 1960; Epstein, 1975; Lindgren & McElratn, 1966; Wald, 1947), and are reproduced in the appendix with the parameters used in the present study.

The objective of this study was to examine the application of sequential analysis in the individual Army training and testing context. Specifically, the effectiveness of sequential analysis was compared to the total score on the same test as used for the classification decision, and to a criterion external to the test itself. In addition, the efficiency of the sequential testing procedure was measured using the percent of the total number of test items that were required to make the sequential decisions. The overall objective then, was to explore the combined effectiveness and efficiency of sequential analysis for the proficiency classification of individuals.

METHOD

Data for 500 enlisted men, at three levels of training on five combat topics, were available from previous research (Knerr, Downey & Kessler, 1975). In that study, hands-on performance tests were administered

approximately a week after the training phase of the research. The topics, total number of items in the tests, and the number of soldiers tested in each topic are shown in Table 1. The topics were Hand Grenades, Light Antitank Weapon (LAW), M16A1 Rifle, Mortar Fire Direction Computer (FDC), and Surveyed Firing Charts.

Scores were available for each soldier on each item and on the total test (i.e., the cumulative number of items correct). An index of training was established, based on the extent of refresher training that the soldiers received during the research. The lowest level (scored 1) was no refresher training, the middle level was some refresher training (scored 2), and the highest level was the most effective refresher training (scored 3) as measured by a criterion other than the tests being used for the present research. On tests administered after the refresher training, soldiers who received the most effective training scored 82%, on the average, while soldiers who received some training scored 55% correct, on the average.

The sequential analysis procedure was applied to the item data, using the parameters described earlier (high cut-off score = 90%, low cut-off score = 70%, alpha = .01, and beta = .10). The items were analyzed in the order that they were administered, so that the classification decisions were those that would have resulted if the sequential procedure were applied during the actual test administration. The classification decision for each examinee, and the number of items required to reach the decision, were noted. In some cases, no decision was made before the total number of test items was exhausted. When no decision was made, the maximum number of test items (total items in the test) was recorded for the examinee. Score codes were 1 for non-proficiency, 2 for no decision, and 3 for proficiency.

RESULTS

The sequential analysis procedure classified 93% of the examinees as either proficient (passed) or non-proficient (failed). No classification decision was made for 7% of the examinees; i.e., when all of their item scores had been plotted, the line remained in the same zone of indecision or indifference. The Mortar FDC test produced sequential decisions in all cases, and the M16A1 Rifle test produced decisions for all but one examinee (Table 2). No decision was reached for between 13-14% of the examinees on the other three performance tests.

The distribution of sequential analysis outcomes (pass, fail, or no decision) among the three levels of training are shown in Figure 4. Averaged over the five tests, 31% of the examinees passed and 62% failed. It is important to note that the tests were administered approximately a week after the training, so that the trained examinees earned lower scores

TABLE

TEST TOPICS, NUMBER OF TEST ITEMS, AND NUMBER OF SOLDIERS TESTED

Test Topic	Number of Test Items	Number of Soldiers Tested
Hand Grenades	29	107
LAW	25	112
M16A1 Rifle	64	117
Mortar FDC	55	105
Surveyed Firing Charts	33	59

TABLE 2

SUMMARY OF SEQUENTIAL ANALYSIS DECISIONS BY TEST TOPIC AND TRAINING GROUP

Test Topic	Sequential Decision	Number of Examinees by Training Group			Total Sample	
		Best	Some	None	No.	%
Hand Grenades	Pass	6	1	0	7	7%
	None	11	2	1	14	13%
	Fail	16	32	38	86	80%
LAW	Pass	9	9	6	24	21%
	None	8	6	0	14	13%
	Fail	20	21	33	74	66%
M16 Rifle	Pass	19	14	7	40	34%
	None	0	1	0	1	1%
	Fail	22	22	32	76	65%
Mortar FDC	Pass	25	13	8	46	44%
	None	0	0	0	0	0
	Fail	10	22	27	59	56%
Surveyed Firing Charts	Pass	18	13	8	39	66%
	None	2	2	4	8	14%
	Fail	1	5	6	12	20%

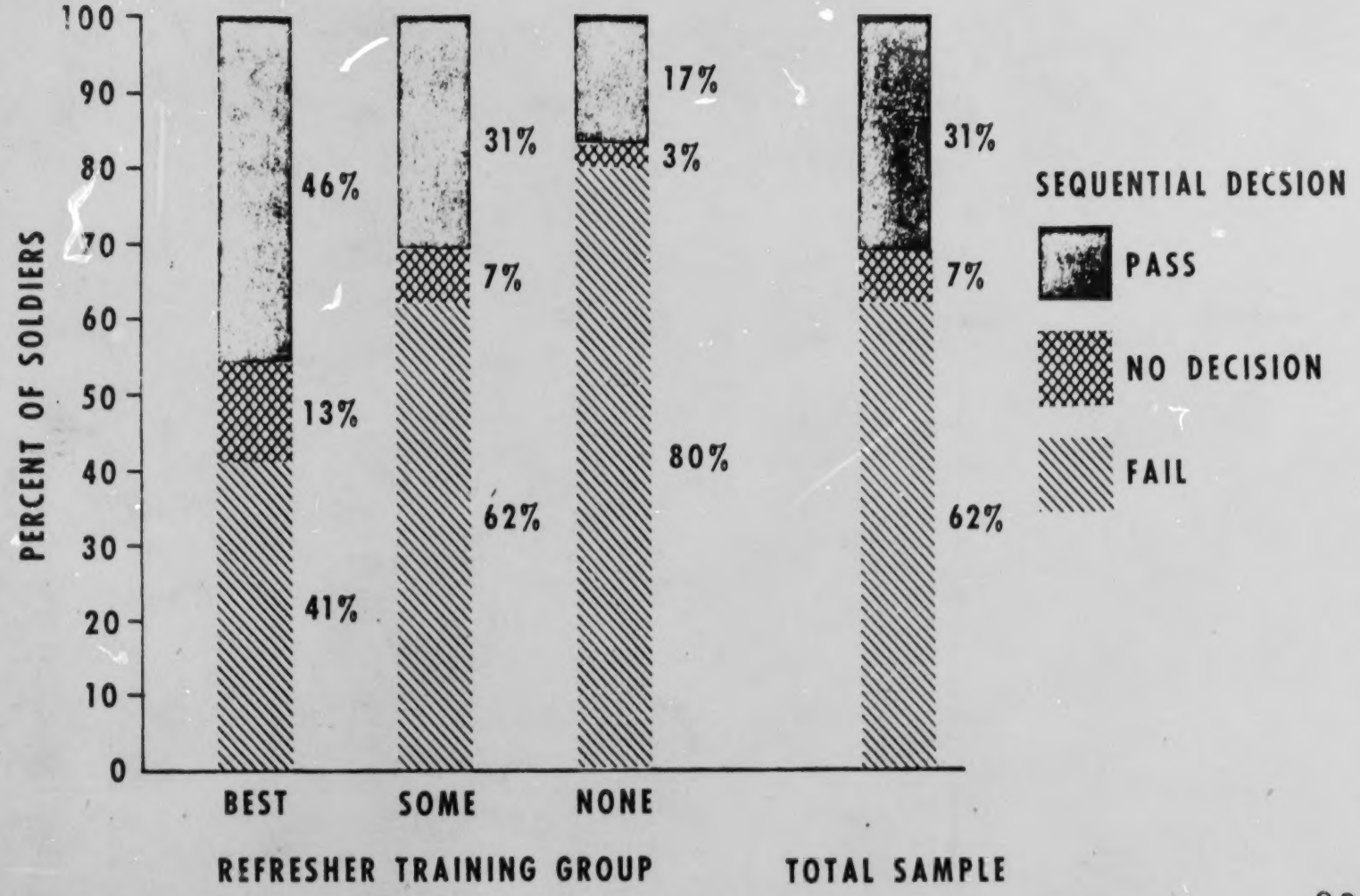


FIGURE 4. SEQUENTIAL DECISION DISTRIBUTION

than the typical 80% correct: the scores immediately after training, reported in the Method section, gave an index of the effectiveness of the training, and were closer to the typical post-training level. Generally, soldiers who received the most effective training more often passed, and soldiers who received no refresher training more often failed. The correlations between the sequential decisions and the training index (Table 3) ranged from .20 (LAW test; $p < .05$) to .47 (Hand Grenade test; $p < .01$), and the average of the correlations was .34. The correlations between the total test scores and the training index ranged from .27 (M16A1 Rifle test; $p < .01$) to .58 (Hand Grenade test; $p < .01$), with an average correlation of .38. Thus, the sequential decisions were about as effective as the total test scores when effectiveness was assessed against a criterion external to the test.

Since decision accuracy, the extent of classification error, is known to be a problem when the classifications are based on a small number of items, the agreement between the sequential decisions and the total test scores (that reflect all of the items) was examined. Where the sequential analysis produced unambiguous proficiency decisions, these decisions agreed with classifications based on the total test score, using the 80% correct criterion, for 88% of the examinees. Thus, the classifications were in error for 12% of the examinees. Examining just the erroneous decisions, 11% were false positives (sequential analysis indicated proficiency, while the total test score indicated failure), and 1% were false negatives (sequential analysis indicated non-proficiency, while the total test score indicated that the examinee passed). These error rates correspond closely to the predetermined allowable error rates, alpha and beta. The value of alpha was set at .01, and the extent of false negative decision error was restricted to this amount in these data. Beta, set at .10, was closely approximated by the obtained error rate for false positives, 11%. Thus, in empirical data, on a sample of 500 soldiers, the observed error rates were close to the predetermined ones, indicating that the sequential procedure functioned properly with regard to control over the degree of classification error.

The efficiency of the sequential procedure was examined by comparing the total number of items in each test with the percent of the total number of items that were required to reach the sequential decisions (Figure 5 and Table 4). Averaged over individual examinees, between 19% (Mortar FDC test) and 66% (LAW test) of the total number of test items were required for the sequential decisions. Over the five performance tests, an average of 33% of the items were required to make the sequential decisions. Thus, two-thirds of the items, test administration time, and cost could have been saved by using the sequential analysis procedure rather than the traditional procedure of administering the entire test.

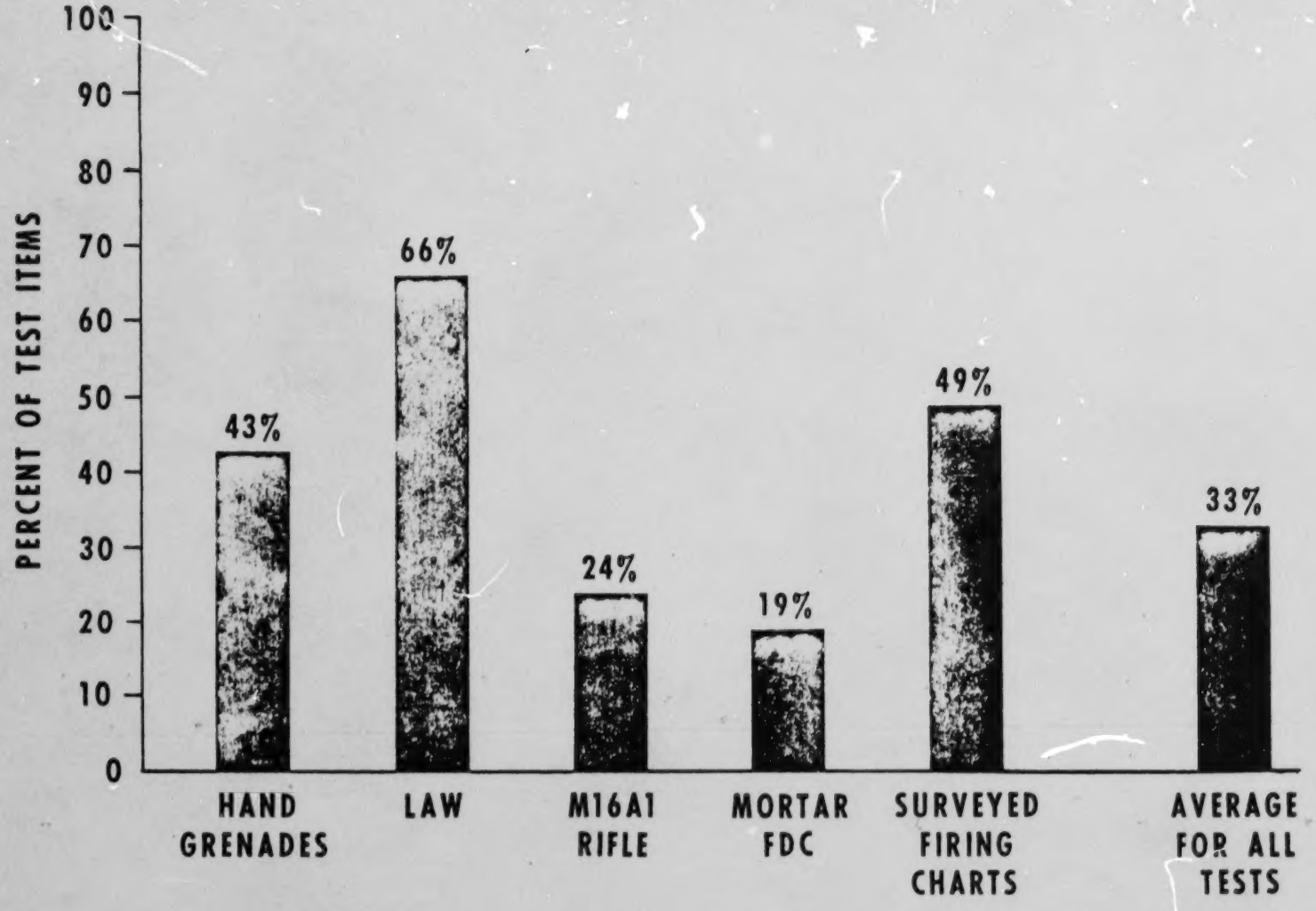
TABLE 3

INTERCORRELATIONS, MEANS, AND STANDARD DEVIATIONS OF SEQUENTIAL DECISIONS, TOTAL TEST SCORES, AND TRAINING INDEX

Test Topic	Correlation with Training Index		Means and Standard Deviations					
			Sequential Decision		Total Test Score		Training Index	
	Sequential Decision	Total Test Score	Mean	SD	Mean	SD	Mean	SD
Hand Grenades	.47**	.58**	1.26	0.57	15.04	6.10	2.06	0.82
LAW	.20*	.34**	1.55	0.83	14.44	4.86	2.02	0.83
M16A1 Rifle	.25**	.27**	1.69	0.95	41.85	11.75	1.98	0.83
Mortar FDC	.40**	.39**	1.88	1.00	21.48	21.73	2.00	0.82
Surveyed Firing Charts	.35**	.28*	2.46	0.82	25.44	8.55	1.95	0.81

** $p < .01$

* $p < .05$



IGURE 5. PERCENT OF TEST ITEMS FOR SEQUENTIAL DECISIONS

TABLE 4

ITEMS REQUIRED FOR SEQUENTIAL

DECISIONS AND TEST INTERNAL CONSISTENCY

Test Topic	Test Items Required for Sequential Decision		Internal Consistency KR20
	Average Number	Percent of Total	
Hand Grenades	12.51	43%	.86
LAW	16.41	66%	.83
M16A1 Rifle	15.68	24%	.93
Mortar FDC	10.43	19%	.99
Surveyed Firing Charts	16.02	49%	.96

Combining the accuracy and efficiency results, it appears that little accuracy is gained by administering the total test rather than the smaller portion required for the sequential decisions.

These data also demonstrated the influence of test homogeneity, or internal consistency, on the sequential decision outcomes. The mathematical basis of the sequential analysis method assumes that the observations are randomly sampled from a single domain. In the present application, the items in a test should all measure the same skill. The order of item administration should be irrelevant, because for any item, passing or failing may indicate the classification decision for the total test. That is, the pass or fail score on any item should correspond to the pass or fail classification decision. This correspondence implies internal consistency, as measured by the Kuder and Richardson (1937) formula 20, for which the results are reported in Table 4. The Mortar FDC test had the highest internal consistency (.99), and the sequential analysis procedure functioned very well in the Mortar FDC data. It classified all of the soldiers, needed only 19% of the total test to reach the decisions, and the sequential decisions were highly related to the training index. Thus, the sequential procedure was accurate and efficient in data that met the homogeneity assumption. In contrast, the LAW test had lower internal consistency (.83), and the sequential procedure did not work as well in the LAW data. More items were required to reach the decisions (66%), a high number of soldiers had no sequential decision (13%), and the decisions did not relate highly with the training index. In short, when the items did not measure the same underlying skill, or did not measure it reliably, then a higher portion of the items was required to reach the decisions, and no decision was reached for a higher portion of examinees. The sequential analysis procedure was more effective for tests that measured a single objective, and comparatively less effective for tests that measured diverse objectives, or were less reliable. If the scores have low internal consistency for any reason, the order of presentation of the items may have an effect on the decisions. This violates the sequential testing assumption that the observations are a random sample drawn from the same domain.

DISCUSSION

The sequential analysis procedure was compared to the traditional testing procedure (administering the entire test before making a proficiency decision) with respect to accuracy and efficiency in dichotomous proficiency classification. The sequential procedure classified examinees about as accurately as the conventional procedure, when accuracy was determined by a criterion external to the test itself. This research replicated the finding that sequential analysis requires considerably fewer test items to produce classification decisions than does traditional testing. The best estimate

of the item savings accrued by sequential testing lies between one-half of the total test items, demonstrated by Linn, Rock, and Cleary (1972), and two-thirds, demonstrated in the present study. The cost for the present performance test administration was over $21,000, so that reducing the testing by two-thirds could have saved as much as $14,000 in this study.

The simplicity of applying the sequential procedure was shown in Figures 1, 2, and 3. Templates can be constructed so that the test administrator marks the examinees' scores during the test and terminates the test when the examinee passes or fails.

The claim that the sequential analysis procedure controls the extent of decision error within predetermined bounds was supported in these empirical data. The obtained error rates closely approximated the limits (alpha and beta) set for the analysis. The sequential procedure produced no more false negative decisions, and only slightly more false positive decisions, than the allowable decision error limits set for the analysis.

Some evidence was presented that sequential analysis functioned better when the assumption of test homogeneity was met than when it was violated. However, this research was not designed to explore the characteristics of the sequential procedure in detail, and results concerning the mathematical assumptions are only suggestive until further research is conducted.

Another future research area concerns the examinees for whom no sequential decision is reached before the item pool is exhausted. Wald (1947) proved mathematically that the sequential test will terminate, and others (see Epstein, 1975) have introduced methods to prevent the need for excessive numbers of items. Validation of these methods, and determination of the conditions that influence them, remain to be explored.

REFERENCES

Crow, E. L., Davis, F. A., and Maxfield, M.W. Statistics Manual. New York: Dover Publications, Inc. 1960.

Epstein, K. I. Sequential plans and formative evaluation. Paper presented at AERA annual meeting. Washington, D.C., April 1975.

Jeter, C. C. and Singleton, J. M. H. Procedures for developing sequential validation. U.S. Army Signal School, Ft. Gordon, GA, July 1975.

Knerr, C. S., Downey, R. G., and Kessler, J. J. Training individuals in Army units: Comparative effectiveness of Selected TEC Lessons and Conventional Methods. ARI Research Report 1188. December 1975.

Kriewall, T. Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Technical Report 103. The Wisconsin Research and Development Center for Cognitive Learning, Madison, Wisconsin, October 1969.

Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 95-101.

Lindgren, B. W. and McElrath, G. W. Introduction to probability and statistics. New York: The Macmillan Company, 1966.

Linn, R. L., Rock, D. A., and Cleary, T. A. Sequential testing for dichotomous decisions. Educational and Psychological Measurement, 1972, 32, 85-95.

Novick, M. R. and Lewis, C. Prescribing Test Length for Criterion-Referenced Measurement. ACT Technical Bull. No. 18, January 1974. Also in C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in Evaluation, No. 3 Los Angeles: Center for the Study of Evaluation, University of California, 1974.

U.S. Army Training and Doctrine Command. TRADOC Pamphlet 350-30, Interservice Procedures for Instructional Systems Development. Fort Monroe, VA, 1 August 1975.

Wald, A. Sequential analysis. New York: Wiley, 1947.

APPENDIX

The sequential likelihood ratio test is designed for testing between two simple hypotheses which, in the case of a Bernoulli population, can be written: $H_1 : p = p_1$

$H_2 : p = p_2$

where p_1 = 1 - the high cut-off score, and

p_2 = 1 - the low cut-off score.

The test is based on the value of the likelihood ratio computed after each observation, including all of the observations obtained up to that point:

(1) $$\lambda_n = \frac{L(p_2)}{L(p_1)} = \frac{p_2^k (1-p_2)^{n-k}}{p_1^k (1-p_1)^{n-k}}$$

where λn = the likelihood ratio

$L(p_2) = p_2^k (1 - p_2)^{n-k}$ = the likelihood of p_2

$L(p_1) = p_1^k (1 - p_1)^{n-k}$ = the likelihood of p_1

and n = the number of observations, and

k = the number of failures.

Two values, A and B are chosen such that if (1) $\lambda_n < A$, then hypothesis $H_1 : p = p_1$ is chosen, (2) $\lambda_n > B$, then hypothesis $H_2 : p = p_2$ is chosen, and (3) $A < \lambda_n < B$, then another item is chosen and another observation is made. Good approximations of A and B in terms of the two types of decision making error, α and β are specified as

(2) $A = \beta/(1-\alpha)$

$B = (1-\beta)/\alpha$.

Combining equations (1) and (2) the inequality above may be written:

(3) $$\frac{\beta}{(1-\alpha)} < \frac{p_2^k (1-p_2)^{n-k}}{p_1^k (1-p_1)^{n-k}} < \frac{(1-\beta)}{\alpha} .$$

In order to simplify the computations, we take the logarithm of each term in (3) and rearrange terms so that (3) is expressed in terms of k, the number of failures.

(4) $$\log\left(\frac{\beta}{1-\alpha}\right) < k \log\left(\frac{p_2}{p_1}\right) + (n-k) \log\left(\frac{1-p_2}{1-p_1}\right) < \log\left(\frac{1-\beta}{\alpha}\right)$$

(5) $$\log\left(\frac{\beta}{1-\alpha}\right) < k \log\left(\frac{p_2}{p_1}\right) + n \log\left(\frac{1-p_2}{1-p_1}\right) - k \log\left(\frac{1-p_2}{1-p_1}\right) < \log\left(\frac{1-\beta}{\alpha}\right)$$

(6) $$\log\left(\frac{\beta}{1-\alpha}\right) - n \log\left(\frac{1-p_2}{1-p_1}\right) < k\left[\log\left(\frac{p_2}{p_1}\right) - \log\left(\frac{1-p_2}{1-p_1}\right)\right] < \log\left(\frac{1-\beta}{\alpha}\right) - n \log\left(\frac{1-p_2}{1-p_1}\right)$$

(7) $$\frac{\log\left(\frac{\beta}{1-\alpha}\right) - n \log\left(\frac{1-p_2}{1-p_1}\right)}{\log\left(\frac{p_2}{p_1}\right) - \log\left(\frac{1-p_2}{1-p_1}\right)} < k < \frac{\log\left(\frac{1-\beta}{\alpha}\right) - n \log\left(\frac{1-p_2}{1-p_1}\right)}{\log\left(\frac{p_2}{p_1}\right) - \log\left(\frac{1-p_2}{1-p_1}\right)}$$

Finally, we take advantage of the fact that the extremes of the inequality in (7) are linear functions of n, compute them, and graph them at the

beginning of the test. The value of k is then plotted as the test proceeds, and when it crosses one of the two straight lines, the test stops and a decision is arrived at.

Simple computational equations for the two straight lines are derived as follows:

$$(8)\quad d_1 = \frac{\log\left(\frac{\beta}{1-\alpha}\right) - n\log\left(\frac{1-p_2}{1-p_1}\right)}{\log\left(\frac{p_2}{p_1}\right) - \log\left(\frac{1-p_2}{1-p_1}\right)} \qquad \text{(left extreme (7))}$$

$$d_2 = \frac{\log\left(\frac{1-\beta}{\alpha}\right) - n\log\left(\frac{1-p_2}{1-p_1}\right)}{\log\left(\frac{p_2}{p_1}\right) - \log\left(\frac{1-p_2}{1-p_1}\right)} \qquad \text{(right extreme (7))}$$

$$(9)\quad g_1 = \log\left(\frac{p_2}{p_1}\right) \qquad g_2 = -\log\left(\frac{1-p_2}{1-p_1}\right)$$

$$a = \log\left(\frac{1-\beta}{\alpha}\right) \qquad b = -\log\left(\frac{\beta}{1-\alpha}\right)$$

$$h_1 = b/(g_1 + g_2) \qquad h_2 = a/(g_1 + g_2) \qquad s = g_2/(g_1 + g_2)$$

$$(10)\quad d_1 = -h_1 + sn \qquad \text{(lower line)}$$

$$d_2 = h_2 + sn \qquad \text{(upper line)}$$

SAMPLE CALCULATIONS

Cut-off scores: high = 90%, low = 70%

Necessary Parameters: $p_1 = .10$ $p_2 = .30$ $\alpha = .01$ $\beta = .10$

Calculations: $g_1 = \log p_2/p_1 = \log .30/.10 = \log 3 = .477$

$$g_2 = -\log \frac{1-p_2}{1-p_1} = -\log \frac{.70}{.90} = -\log (.7778) = .109$$

$$a = \log (1-\beta)/\alpha = \log .90/.01 = \log 90 = 1.954$$

$$b = -\log \frac{\beta}{1-\alpha} = -\log \frac{.10}{.99} = -\log (.1010) = .996$$

$$h_1 = b/(g_1 + g_2) = .996/ (.477 + .109) = .996/.586 = 1.70$$

$$h_2 = a/(g_1 + g_2) = 1.954/(.477 + .109) = 1.954/.586 = 3.33$$

$$s = g_2/(g_1 + g_2) = .109/(.477 + .109) = .109/ .586 = .186$$

$$d_1 = -h_1 + sn = -1.70 + .186n$$

$$d_2 = h_2 + sn = 3.33 + .186\ n$$

INTRODUCTION TO THE SYMPOSIUM

CLAY V. BRITTAIN
US ARMY TRAINING EVALUATION GROUP

In a sense, this symposium is a return engagement. Last year, in Indianapolis, the MTA program included a session on skill qualification testing, but it seemed appropriate for us to request program time for discussion of the skill qualification test (SQT) program again this year. Much has happened during the past twelve months.

SQT in four military occupational specialties (MOS) have been developed, printed, and distributed for Army-wide administration during the first quarter of FY 1977 (October-November 1976). This is a field trial designed to shake down the SQT system. Results of these tests are to be used only for advisory/informational purposes. The first SQT for official use will be conducted in the third quarter of FY 1977 (April-June 1977). Soldiers in seven combat arms specialties will be tested at that time. Thereafter, SQT in various MOS will be phased into the program until the full range of SQT are operational some three years from now. The Individual Training Evaluation Group (ITEG) has been established at Fort Eustis, Virginia, and been given responsibility for administering the SQT program.

SQT are developed by test development agencies (TDA), which are the Army schools having proponency for the various enlisted MOS. The proponent school for an MOS is responsible for developing the SQT in that MOS. The test is developed under ITEG's technical supervision and in accordance with guidelines issued by ITEG. The SQT development and validation cycle can be briefly described as follows:

Phase I - The TDA formulates the test plan and submits it to ITEG for review. ITEG approval of test plan marks the end of Phase I.

Phase II - The TDA develops sample units of the SQT and submits them to ITEG for review. Following ITEG approval of these sample units, the TDA develops the complete SQT and tries it out locally. The SQT, along with the results of the tryout, are submitted to ITEG for review. ITEG approval of the draft SQT marks the end of Phase II.

Phase III - The TDA incorporates any required changes in the SQT and then gives it a field tryout. Following this field validation, the TDA makes changes as necessary and submits the test to ITEG for final review and approval.

As will be clear from the papers which follow, skill qualification testing represents a markedly different approach to testing the MOS proficiency of soldiers. It may be useful in setting the context for the symposium to note that the SQT program has emerged and is evolving within a broader context of change in the Army. SQT were approved in concept by the Army Chief of Staff in August 1974 as a part of a new enlisted personnel management system (EPMS). EPMS involves the restructuring of MOS and the realignment of enlisted career management fields to give soldiers more equitable and coherent patterns of career progression, the establishment of more systematic relationships of pay grades to skill levels, and the modification of the NCO education system to give a better fit between training and job requirements. The SQT also is consistent with emphasis in the Army Training and Doctrine Command that training be based upon rigorous job and task analyses and that more of the responsibility of training be shifted from the schoolhouse to the field unit and the individual soldier.

In this symposium, we will describe SQT and discuss their development and validation. In doing so, we feel that we are not simply talking about SQT but about some basic issues in occupational proficiency testing.

The SQT Program as a Criterion-Referenced Approach to Proficiency Testing

Arthur G. Hermansen, US Army Individual Training Evaluation Group

Testing to evaluate the proficiency of an individual in his occupation is not a new concept. A Chinese emperor of antiquity, Emperor Shun by name, used to test his officers every three years. After three such tests, he would either promote his officers or dismiss them from his service. His primary reason for this evaluation was to provide men of ability for the state. The road to test development has been, in terms of time, a long and difficult venture. The road has been chuck-holed with multiple choice tests, washed out by power tests, detoured by tedious essay tests, broken by interest tests, eroded by aptitude tests and closed by the sign posts of jurisprudence. Even in the face of such a formidable enterprise, a new construction outfit has accepted a contract to provide men of ability for the service of the state. This organization is the US Army Individual Training Evaluation Group, a part of the Training Support Center at Fort Eustis, VA. The new road gang is proceeding under the assumption that some technological progress in the area of testing has been accomplished since the rule of the Great Shun, and we really ought to take advantage of these concepts and innovations where appropriate.

One of the popular ideas going strong in testing and other philosophical groups is the criterion-referenced (CR) measurement concept. In fact, CR has become a national buzz-word. I have heard of criterion-referenced meat loaf and criterion-referenced conjugal bliss. Incidentally, both items are performanced based. The criterion-referenced concept has been with us for years waxing then waning in the professional literature. As early as 1924 a statistician, by the name of J. P. Gillford, devoted a chapter of his statistical text to the folly of criterion-referenced tests. For the past 40 years, articles have appeared and are appearing in the professional journals both advocating and criticizing the application of criterion-referenced measurements. No one is really putting down the concept; the difficulty is in trying to force criterion-referenced data into norm-referenced ideas. A working level definition of criterion-referenced measures would go something like this. Criterion-referenced measures are those which are used to determine an individual's status with respect to some criterion, for instance, a specific performance standard. It is because an individual is compared with some criterion rather than other individuals that these measures are described as criterion-referenced. Can the individual accomplish the task or can he not accomplish the task where the task, suitably defined, is the criterion? If we imagine a continuum where one end of the continuum represents no skill whatsoever, in relation to the selected task, and the other end represents superlative skill you would find, in a norm-referenced measurement, individual points along the line representing achievement scores of individuals, an indication of that individual's score in relation to the other individual's scores. Visualizing the same continuum in a criterion-refererenced situation you would find one point

on that continuum established by experts as to the skill results required to adequately express mastery of the task. All scores relate to accomplishment of the task not the other individuals. Borrowing from a Popham and Husk article, we can illustrate by using the criterion-referenced Red Cross Senior Lifesaving test where, "the individual must display certain swimming skills to pass the examination, irrespective of how well others perform on the test." The basic assumption is that retention and career progression of soldiers should be based upon demonstrated performance of critical job related skills. Critical skills then become the criterion to be referenced for testing. It seems reasonable then, to call this new test a Skill Qualification Test and to try to test through multiple instruments those critical skills a soldier needs to do his job. If we do not test for critical skills, a future MTA meeting may be conducted in a foreign language.

First we have to find out what their skills are. Now anyone who has seriously undertaken a job/task analysis can reflect upon the very large number of tasks and subtasks, elements and subelements associated with a specific job. Not all tasks are critical, but it is difficult, in some cases, to determine criticality. Should you decide upon one of several existing methods to determine task criticality, you will find after paring down the list, you still have a rather large domain of critical tasks. In order to test this large domain, you would have to use up several days for testing and significant resources. Most military administrators simply cannot spare their men or equipment for such a time frame. The only recourse is to sample from that domain of critical tasks and try to cover as much of the domain in sequential years testing as you can. The domain we have talked about is not a static pile of critical tasks. It is as if it were a living dynamic force that changes shape and direction as technical innovations, tactics, modifications, and major systems changes take place. The critical skills need constant review and updating.

How do you test these critical tasks? Our answer is to devise scorable units for each of three modes or components of testing.

Scorable units (SU) are administrative devices that link together tasks selected from the domain of critical tasks or relatively large subtasks and a specified number of items. SU are related to specific tasks.

The three components of the SQT include the written component, the hands-on component, and the performance certification component. The written component includes written performance and/or performance based questions. The difference between these items is that:

a. A written performance SU is one where the soldier performs by using the tools and materials or adequate simulation of the written component station. Illustrations of these written performance items are: calculating the distance along a route or map or determining range to target using the mil formula.

b. A performance based SU measures the examinee's ability to answer questions about <u>how</u> he would perform a task. For example, an examinee might be asked to distinguish, before operation maintenance tasks, on a vehicle from a larger list of other maintenance tasks.

Usually a written scorable unit will be a mixture of written performance and performance-based items. The written scorable unit can include up to 10 items. Within each item, there are up to 10 answer spaces. Multiple correct answers are possible. The answers and the distractors <u>must</u> be real-life possibilities. In the written component, it is possible to have 50 scorable units. We ask the test development agencies for a minimum of 30 written component scorable units.

The second part of the SQT is the performance certification component (PCC). Despite some weaknesses of the PCC as a means of measuring a soldier's proficiency, it appears to be the only way that certain highly skilled performances can be evaluated. Performance certification is the component where we consider testing a task that (a) cannot be validly tested in the WC because of its skilled hands-on requirements, and (b) cannot be tested in the HOC because of administrative constraints. Sometimes a hands-on skill simply takes too long to demonstrate, or requires equipment, terrain, or other conditions that prevent its testing in the HOC. Cutting a roadbed with a bulldozer is a good example. To test this task in the HOC would probably take several hours and a sizeable expanse of terrain; moreover, equipment and gasoline requirements could probably not be justified merely for testing purposes.

The thinking behind the PCC is that such tasks can be "tested" on the job. By moving the test to the job site, we can have the soldier's supervisor serve as the scorer in evaluating the man's task performance during the normal course of his job duties. Some natural PCC already exist. Annual rifle qualification is one such example, where a soldier's task proficiency is evaluated in the field, but under conditions essentially the same as one would set up for a hands-on test. And this task is being included in the PCC for relevant MOS. Most tasks that are candidates for the PCC, however, will <u>not</u> already be structured for testing in a field unit, so the test developer must prepare a scoresheet and carefully phrased instructions for the evaluator (supervisor) on how, when, where, and under what conditions a soldier's task performance is to be scored.

Weaknesses inherent in the PCC should be carefully noted. These weaknesses are of two general types - both involving problems of standardization. The first pertains to difficulties in achieving standard conditions for task performance for all soldiers. In setting up a test, as is done in the HOC, we took great care in insuring that conditions were the same for each soldier going through the test station. In the PCC, where a soldier's task performance will be scored as part of his normal job duties, it will be nearly impossible to control the equality of conditions from man to man and location to location. A second weakness in the PCC is that the scorer (supervisor) who is beyond control of

a test control officer, in most cases will not be as well trained or as motivated to do a conscientious and unbiased job of scoring as a scorer in the hands-on component.

The third component, the hands-on component, is probably the most innovative part of the SQT. The aim of this component is to address combat critical tasks that the soldier must know. He must demonstrate his mastery of these tasks by doing. Examples of such tasks are emplacing a mine, loading and firing a machinegun, delivering a grenade on a target, and setting up a field radio. It is probably the most difficult component to administer. Field testing sites must be standardized as much as possible all over the world. Administrative instructions must be developed for all soldiers acting as scorers, and scorers must be trained. For instance, when a grenade lands on the target circle is it in or out?

Most important about the SQT, is the fact that we do not want to keep the test a secret. We want the soldier to know what kind of test he is going to take, what the standards and conditions are, and how to pass the test. Six months before the soldier takes the test he will be furnished a soldier's manual that gives information about the critical tasks he must know, calls attention to important steps or operations in that critical task, and furnishes information as to where he can find out more about the critical tasks. Every soldier will have one and as required, he will receive updated information for his individual manual.

Sixty days before the selected test quarter, we will provide the soldier with an SQT notice. This notice will tell the soldier what this test will ask him. For instance, he may be told that in a written component scorable unit he will be asked to locate and calculate azimuths and back azimuths on a specified map or he will be required in a hands-on test to don a gas mask in nine seconds. He will also be told what he needs to know in order to do the critical task on pure criterion-referenced GO NO-GO basis. We want him to practice so that he passes the test. For those of you who think this is more training than testing, you may be right. The managers of the United States Army want to know where training needs to be modified or changed. Field commanders want to know where their soldiers need additional training emphasis, but most of all criterion-referenced testing to realistic criterion will assure us that when the clock strikes 12 we will not turn into pumpkins.

DEVELOPING SKILL QUALIFICATION TESTS

October 1976

Military Testing Association Conference

Gulf State Park, Alabama

William C. Osborn and Roy C. Campbell

Human Resources Research Organization
Fort Knox, Kentucky 40121

DEVELOPING SKILL QUALIFICATION TESTS

William Osborn and Roy C. Campbell

You have some feel at this point for the scope and direction of the SQT program. To develop and administer a performance-oriented test of job proficiency to nearly every soldier in the United States Army is, indeed, an ambitious undertaking. It is awesome enough to contemplate the logistics of test administration, scoring and reporting - but add to that the front-end requirement for test development. Those of you who construct tests I'm sure, can, appreciate the magnitude of the SQT development effort.

Since ITEG is staffed with no more than a handful of test psychologists, they could not possibly hope to construct SQT's in the numbers required. The mission of developing SQT's therefore resides with the Test Development Agecies -- that is, TRADOC centers and schools which have doctrinal and training responsibility for the various categories of Army jobs. ITEG, then -- as indicated by Mr. Hermansen -- serves as planner, coordinator, and quality control agent for the SQT system. One of their chief responsibilities in this role is to provide detailed guidance to the TDA's on procedures for SQT development. The substance of these procedures and methods for their implementation are what I would like to describe for you today. First I will give an overview of the SQT development process; then I will discuss in more detail what we believe are two of the more interesting and challenging steps in the process; and finally, I will describe briefly what ITEG is doing to support the TDAs in their development work.

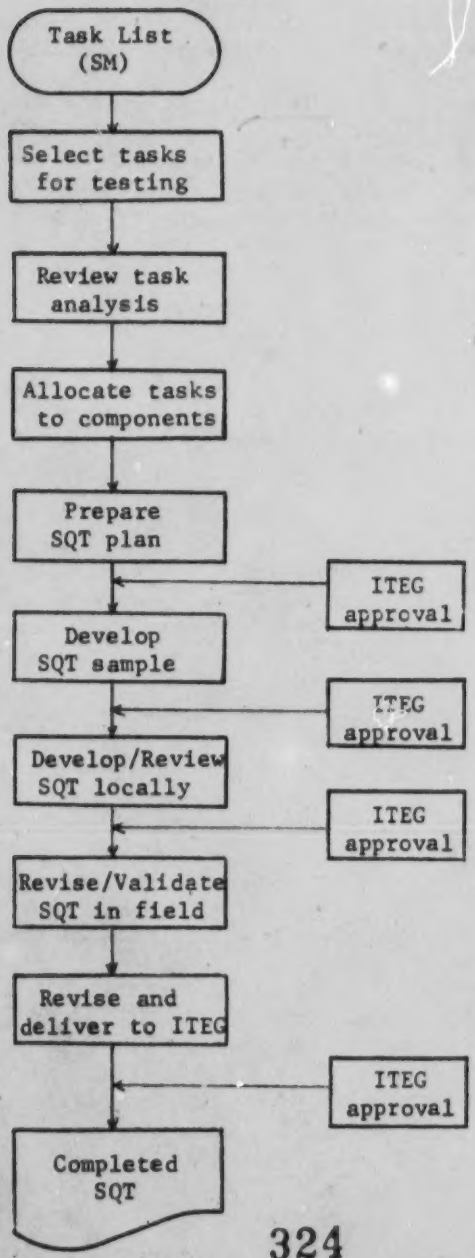
Development Process

This chart gives an overview of the development process. Keep in mind that the steps shown here are performed by the TDA with ITEG review and approval.

Task List. The development process begins with the tasks from the Soldier's Manual. You will recall that the Soldier's Manual is the pivotal document in the Enlisted Personnel Management System. It describes for the soldier exactly what he or she must know and be able to do on the job; it is comprehensive, detailed, and, above all, written in performance terms. All individual training and testing proceeds from this manual, and the SQT is no exception.

Select Tasks for Testing. Since the task list in a Soldier's Manual is comprehensive, and since test administration time is limited, not all tasks can be tested. The first step in SQT development, thus is to identify the more important job tasks. Emphasis is placed on task importance in sampling rather than representativeness by functional areas. Task importance is defined by two factors: (a) criticality to mission accomplishment, and (b) performance deficiency. All tasks in the Soldier's Manual are important, of course, but for purposes of testing some are more important than others. In sorting them out, the developer draws on a variety of data sources, such as tactical doctrine, results of field unit tests, IG inspection reports, maintenance deficiency records, and the judgments of experts. These data are processed and combined in providing at least a roughly ordered task list from which a managable number can be selected for testing.

OVERVIEW OF SQT DEVELOPMENT PROCESS



Review Task Analysis. Having selected the tasks to be included in the SQT, the next step is to review the task analyses. Good tests can proceed only from good task detailing -- a point that cannot be over-emphasized. Because of limitations on space in the Soldier's Manual, task descriptions are sometimes abbreviated. Each task must therefore be thoroughly inspected by the test developer and a subject matter expert to verify the completeness and accuracy of task elements, task conditions and standards. Also during the review process the developer identifies what are termed "key" task elements. Key elements are identified using criteria similar to those used in identifying important tasks. A key element is a subtask or step which is frequently performed incorrectly, or which, if performed incorrectly, results in serious injury to the soldier or damage to expensive equipment. This information is used later in test development.

Allocate Tasks to Component. The next step in development is to decide which of the three SQT components -- Hands-On, Written or Performance Certification -- a task will be tested in. Allocation decisions are based on consideration of task content and the strengths and weaknesses of the components as modes of testing. I will return in a moment to this topic since it is one of the more subtle and challenging steps in SQT development.

Develop SQT Plan. Work performed up to this point by the test developer is documented in an SQT Plan and submitted to ITEG for review. The plan has two parts: (1) a list of tasks to be tested by SQT component, and (2) an explanation of the rationale underlying task selection and allocation.

Develop SQT Sample. Once the SQT plan has been approved, the Test Development Agency gets down to the substance of actual Scoreable Unit development. A Scoreable Unit, you will recall, is that part of an SQT that taps performance of a job task or subtask. A Scoreable Unit in the Hands-On Component consists of four things: (1) standard specifications for the test station at which soldier performance will be observed, (2) instructions to be read to the soldier, (3) performance measures to be observed and evaluated by the scorer, and (4) the pass-fail (or "GO/NO-GO") scoring standard. A Scoreable Unit in the Written Component consists of from one to ten multiple-choice questions about task performance. In preparing items for a Written Scoreable Unit the developer is careful to base questions on key task elements, to elicit a test behavior from the examinee that is as close as possible to task behavior, and to offer only real-world response alternatives. A Scoreable Unit in the Performance Certification Component is similar in substance to those in the Hands-On, but is designed for use by a soldier's supervisor in systematically evaluating task performance in the actual job setting. Since procedures for SQT development are quite different from those used in the past, ITEG finds it efficient to take a look at a few Scoreable Units before the TDA gets too far along in the development process. Therefore, they request for each SQT a sample consisting of five Written Scoreable Units, two Hands-On Scoreable Units, and all of the Performance Certification Component. They also ask to see the supporting task analyses, special administrative instructions, and draft entries for the SQT Notice.

Develop/Review SQT Locally. When ITEG approves the sample Scoreable Units, a TDA proceeds to develop the entire SQT. The first stage of

validation is also carried out in the form of local reviews by job experts and small-scale tryouts. Scoreable Units are thus checked for completeness, job relevance, and perceived fairness. Clarity and readability to the soldier are points of emphasis in reviewing Written Scoreable Units; reliability of scoring is the major factor examined for Hands-On SUs. The completed SQT is then shipped to ITEG.

Revise/Validate SQT in Field. Revisions, based on results of ITEG review, are made before the SQT is taken to the field for validation. When this is done, the TDA, in conjunction with ITEG, selects a field unit for tryout of the Hands-On Component. The TDA submits the entire component including administrative instructions to the Test Control Officer for the unit selected. The chief purpose of this validation run is to check the administerability of the Hands-On Component. The Written Component is validated on a group of up to 30 soldiers, approximately half of whom are masters and half non-masters. The objective here is to establish the discriminant validity of each Written Scoreable Unit. The Performance Certification Component undergoes subjective validation at this stage. It is submitted to a sample of commanders and supervisors who assess its relevance and administrative feasibility.

Revise and Deliver to ITEG. Results of the field validation are used by the TDA in preparing a final draft of the SQT. When approved by ITEG, the SQT is ready for operational administration which concludes the development process. Printing and distribution; and later, scoring, data reduction, and distribution of test results are all carried out by ITEG.

Special Development Features

The foregoing overview of SQT development was sketchy at best. Time unfortunately, does not permit a thorough description of each step in the process. It should be mentioned, however, that ITEG has prepared detailed procedures, supported by flow charts, for each step in the sequence described. I would like to give you some feel for this level of detail by discussing two of the more thought provoking aspects of SQT development: allocating tasks to component, and preparing Written Scoreable Units. I believe that in both of these aspects, ITEG is pushing the state-of-the-art in performance test development, and therefore should be of interest.

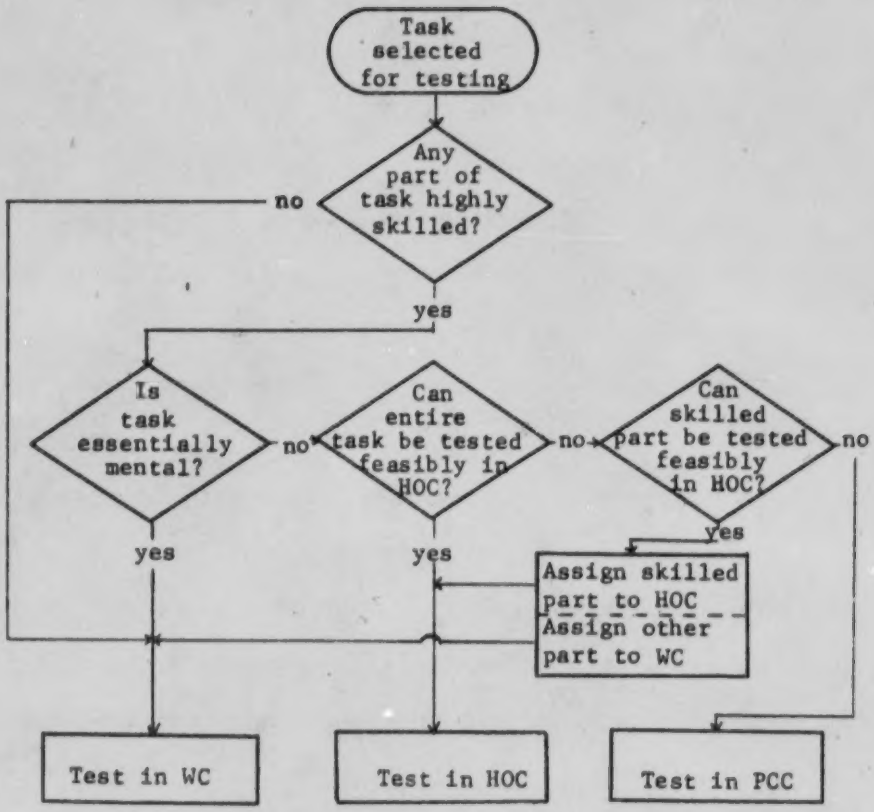
Allocating Tasks to Component. Determining the appropriate component for a task is not easy. Yet every task must be assigned to some component. Guidelines were formulated, therefore, based on the characteristics of job tasks and test components. Criteria pertaining to task content center on the type of task and the degree of skill required in its performance. Two types of tasks are considered, mental and physical, and two levels of skill, low and high. All types and levels of skill can be tested validly in a hands-on mode, assuming unlimited resources. Some physical tasks may be tested in the Written Component, but highly skilled physical tasks must be tested in the Hands-On or Performance Certification component.

The criterion pertaining to test component was derived from testing feasibility. Any testing requires some resources. In the case of SQT, the Written Component is the least demanding since, other than occasional use of job aids and AV media, only paper and pencil are required;

moreover, many soldiers can be tested on many tasks with one test administrator in a relatively short time. The Hands-On Component, on the other hand, requires equipment, terrain, and several qualified scorers; in addition, relatively few tasks can be tested in the four hours available. The Performance Certification Component was designed to accomodate tasks that are extremely time consuming and that require special terrain, facilities, and expensive equipment; it is of course the weakest component from a measurement standpoint.

The chart shows the decision flow recommended for allocating tasks. The first question considered is "Is any part of the task highly skilled?" "Any part" refers to subtasks or task elements. "Highly skilled" should be thought of in terms of how much practice is required to become proficient. Or, looking at it from a different point of view, can the task be performed by someone who knows the procedure even though they may have done it only once or twice? If no part of the task is highly skilled, or if the only skill involved is mental, such as remembering a series of steps, performing a numerical calculation, or understanding and selectively applying a principle, then the task is assigned to the Written Component. Tasks not in this category, that is, highly skilled tasks, include the following: (a) finely tuned acts of physical coordination (e.g., tracking a target with a tank gun), (b) tasks requiring quick reaction or that must be performed with a tight time limit (e.g., clearing a jammed weapon, or typing a message), (c) tasks involving time-sharing (e.g., tactical driving). If the task does involve a physical skill, the next consideration is whether the entire task can be tested feasibly in the Hands-On Component. If time, equipment, terrain and qualified

ALLOCATE TASKS TO SQT COMPONENT



scorers are generally available to support the Scoreable Unit, then the task is earmarked for Hands-On testing. If the entire task cannot be tested feasibly in the Hands-On Component, perhaps one or more subtasks or key elements can be. All highly skilled elements that can be tested practically in the Hands-On mode should be allocated to that component. When a task is partitioned in this way, the nonskilled or knowledge elements are allocated to the Written Component. If neither the entire task nor any skilled elements can be tested feasibly in the Hands-On Component, the task should be assigned to Performance Certification.

The initial allocation of tasks may have to be adjusted somewhat, since there are limits on the size of each component. Different Army jobs have different types and degrees of skill, and tasks probably should not be forced into a prescribed distribution over components. Fortunately most jobs contain few highly-skilled physical tasks, and, since any task can be tested validly in the hands-on mode, some flexibility is available in the allocation process.

Developing Performance-Based Written Items. I would like to show you one more example of the kind of detailed procedural aid that has been prepared for the test developer. The chart you will be looking at outlines the procedure for developing one type of Written Scoreable Unit. There are two tracks the developer may take in preparing items for the Written Component: one we call Written Performance, the other, Performance-Based Written. The former is essentially a performance test administered within the confines of the Written Component. Despite their desirability, relatively few tasks can be translated into

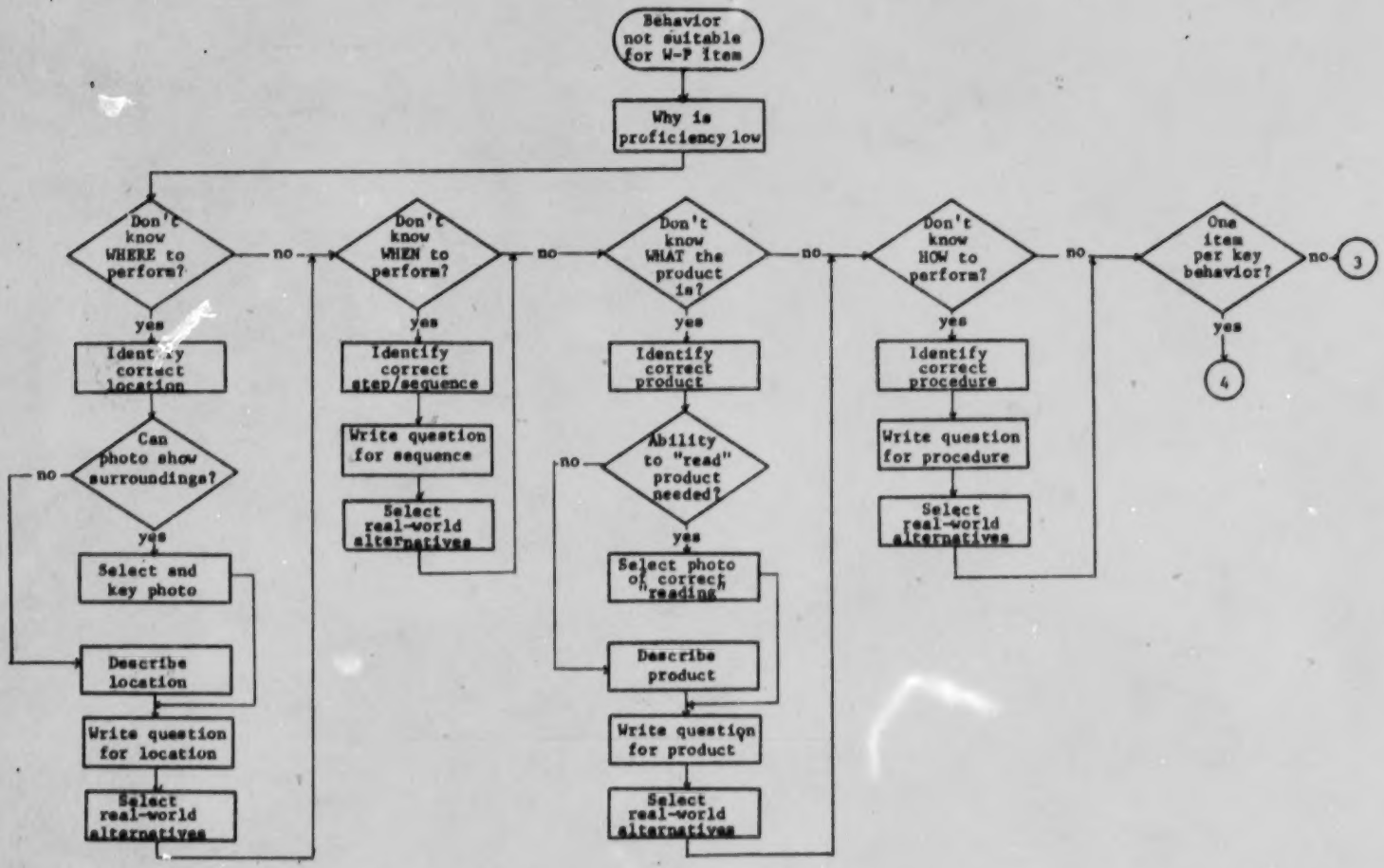
pure Written Performance Scoreable Units. The other class of Written Scoreable Units, referred to as Performance-Based, is the more common. It also is the most similar to conventional, multiple-choice tests of job knowledge. Two features, however, make it worth reporting: one is that test questions are tied intimately to the steps in task performance, and the other is the attempt at proceduralizing development.

Here is the flow chart. It is important to understand that the considerations embodied in this chart are applied not to a task but to the task elements, or what here are termed task behaviors. One proceeds through the chart for each key behavior if not for every behavior in task performance; the object being to prepare a test item for each.

The basic question posed is: Why does the soldier fail to perform this behavior correctly? Subordinate questions are posed in the first four decision diamonds:

> Is it because he doesn't know WHERE to perform? Failure to carry out an aspect of task performance may be caused by not knowing where certain objects are located. (A soldier, for example, may not be able to perform a maintenance check because he does not know where the component is located.)
>
> - or -
>
> Is it because he doesn't know WHEN to perform a step? Failure to perform may be caused by not knowing the correct sequence of steps. (If, for example, the handguards on the M16 rifle are

PREPARE PERFORMANCE BASED ITEMS (3)



not removed at a particular point early in disassembly, the weapon could be damaged.)

- or -

Is it because he doesn't know WHAT the end result looks like? A soldier may know when, where, and even how to perform a step, but doesn't know what the correct outcome is. (In positioning the buzzer volume control knob on the field telephone, for example, a soldier may know where the knob is located, when to turn it, and certainly how to turn it, but doesn't know what the correct position is.)

- or -

Is it because he doesn't know HOW to execute the behavior? The fourth and perhaps most common reason for failure is that the soldier does not know the correct procedure for carrying out some part of the task. (For example, in signaling another station with the field telephone, a soldier may know everything he needs to about when and where to perform, and what the desired result is, but he may not know how to do it.)

When one of these questions is answered affirmatively, the developer follows that branch of the chart in preparing an appropriate test question. Generally, he first identifies the correct location, or sequence, or product,

or procedure; then describes in words or pictures; then frames a question to call it out; and, finally, selects real-world response alternatives -- distractors, if you will -- to complete the test item. I will not talk you through each branch since each follows the same basic line of thought. The important point is that by answering these four questions about each aspect of task performance, the developer is well along the way to pinpointing both what is important to ask in a knowledge test of task performance, and how to ask it. The procedure helps prevent test questions that so often appear merely because they are easy to ask. Moreover it gives a developer some direction; he is less apt to slip into that "Gee-I-wonder-what-I-can-ask-about-this-task" syndrome.

Training Test Developers

Procedural charts of this sort have been prepared for all aspects of SQT development, from constructing Scoreable Units, through validation, to preparing administrative instructions. But since the SQT system departs radically from earlier methods of MOS testing, developers need more than printed guidance; they also need training.

Assisting TDAs in training their personnel is a continuing part of ITEC's program. In the summer of 1975 a preliminary series of workshops, directed at those responsible for the first round of SQTs, were conducted at four TDAs. These workshops were given under the EPMS implementation program and provided early guidance on test development -- particularly for hands-on tests. By the spring of 1976 it was evident that TDAs needed better guidance. ITEC then prepared a Handbook for SQT developers which detailed the procedures that I have outlined for you today. This Handbook served as the basis for a second round of workshops conducted for

the 12 TDAs involved in Phase I of the SQT program. The workshops emphasized practical work, with developers actually producing Scoreable Units. The success of these workshops highlighted the need for an on-going program of training in SQT development. ITEG is responding to this need by converting the handbook and workshop materials to a self-instructional package, covering both development and validation. Only through such a concentrated program of support to Test Development Agencies can the unique concepts of Skill Qualification Testing be realized.

ESTABLISHING THE VALIDITY OF SQT

CLAY V. BRITTAIN
INDIVIDUAL TRAINING EVALUATION GROUP

From the inception of the SQT program, there has been strong emphasis upon the need for test validation. ITEG holds firmly to the position that the validity of an SQT must be established prior to being implemented. Much effort has gone into the development of detailed guidelines to be followed by test development agencies to insure the validity of the SQT. But, my interest here is not so much in specific validation procedures as it is in the general problem of validity as applied to skill qualification testing. That is, the focus of this paper lies in the definitional and conceptual aspects of SQT validity. The positions set forth here are my own. The paper is not a statement of official policy.

In addressing the problem of SQT validity, I find it useful to refer to Lee J. Cronbach's comprehensive chapter on test validation in the second edition of Educational Measurement, American Council on Education, 1971. Cronbach says: "Narrowly considered, validation is the process of examining the accuracy of a specific prediction or inference made from a test score....More broadly, validation examines the soundness of all the interpretations of a test--descriptive and explanatory as well as situation-bound predictions." In this broader sense, SQT validity pertains to the interpretations of SQT scores and is to be determined in terms of the purposes of skill qualification testing.

An SQT serves two general purposes. On the one hand, it provides information relevant to the evaluation and management of individual training. The aim is to identify performance deficiencies which point to training needs and to mobilize and direct training efforts. On the other hand, the SQT provides information for personnel management with regard to promotion, reclassification, and retention.

In terms of its structure, an SQT is based upon tasks which have been identified as critical in an MOS. The SQT consists of modules called scorable units. Each scorable unit reflects a critical task (or task element) and is designed to test the soldier's ability to perform the task (or element). On each scorable unit, the soldier is scored "Go" or "No Go" and receives a score of 1 or 0. These unit scores can be summed to yield task scores, e.g., 40 percent of the soldiers tested scored "No Go" on Task X. This information is useful in mobilizing and directing training effort. The unit scores can also be summed to give the individual soldier an overall SQT score, e.g., Soldier A scored "Go" on 80 percent of the tasks on the SQT. This information is used in personnel management decisions.

It seems to be to be both meaningful and useful to think of SQT validity in terms of three levels:

(1) At the level of the scorable unit, validity pertains to the accuracy with which soldiers are identified as masters or nonmasters of a specified task and to the criticality of the task.

(2) At the level of task scores, validity pertains to performance deficiencies and training effectiveness. Do SQT results reflect real-world performance deficiencies? Is the SQT instrumental in bringing training efforts to bear in improving real-world performance?

(3) At the level of overall SQT scores of individual examinees, validity pertains to the soundness of career-related personnel decisions about soldiers.

I wish to examine validity issues at all three levels but to treat, in greatest depth, scorable-unit validity. The validation of scorable units is basic to validity at the other two levels.

The evidence which is required to validate scorable units is, in certain important respects, different for the different SQT components. With regard to the written components, especially the performance-based scorable units, evidence of concurrent validity is important. That is, when soldiers are identified as "GO" or"NO GO" based upon scorable unit performance, are the identifications consistent with their identification as masters or nonmasters by some other credible source? The problem, of course, is in finding that other credible source for identifying masters and nonmasters.

This is not a problem that we have adequately resolved to date. As an expediency we have asked soldiers to rate themselves in terms of ability to perform the referenced task and have examined the "GO"/"NO GO" scores in terms of consistencies with the self-ratings. We take it that inconsistencies here at least tell us that a scorable unit needs to be thoroughly reexamined.

Another approach would involve the use of hands-on scorable units to provide criterion measures against which to validate written scorable units. TDA have been directed to select a limited number of critical tasks allocated to the written component and to develop hands-on tests for these selected tasks. This appears to be a logical and potentially useful approach to the validation of written scorable units.

In validating scorable units in the hands-on component, the difficulty of scoring soldiers' performance objectively and consistently is of major concern. In validating hands-on scorable units, we must deal

with this problem in terms of interscorer agreement. The TDA must insure that the scoring criteria and scorer instructions are specified clearly enough to yield a high level of interscorer agreement.

Validation problems associated with the hands-on component appear in more extreme form in the performance certification component (PCC). Of special concern is the possibility of rater leniency which might be so pronounced as to seriously impair the validity of the PCC. This is a matter to which we plan to give special attention.

Now, having noted that the requirements for validating a scorable unit are different, depending upon whether it is in written, hands-on, or performance certification mode, we should point out that certain validity issues cut across SQT components. In addressing the scorable unit validation requirements which are common to the different components, it is useful to draw again upon Cronbach's chapter on test validation. Cronbach notes that generally the use of test scores falls into either of two categories; i.e., a test score can be used to make a decision about the person tested or to describe him. This, of course, is not a neat and precise distinction. The use of a test score to make a decision about an individual inescapably involves labelling him; i.e., describing him in at least a limited way. And, of course, descriptions of an individual based upon a test score may enter significantly into decisions made about him. Thus, the distinction between test use for making decisions about individuals and for describing them is approximate and relative, rather than precise and absolute, but, it is nonetheless a meaningful one which has important implications regarding the validation of scorable units.

The scoring of a soldier as "GO" or "NO GO" on a scorable unit is more a matter of describing him than of making a decision about him, and much of what Cronbach says about validation of test scores used descriptively would seem to apply; i.e., validity involves mainly questions of educational importance (rephrased here as task criticality), content validity, and construct validity. These requirements apply to scorable unit validation in all of the components.

Task Criticality: Evidence of task criticality involves:

(1) Task performance deficiencies.

(2) Contribution of the task to critical combat systems (need for task in ARTEP).

Performance deficiencies are identified based upon the judgment of senior job incumbents and supervisors and also on the basis of harder data. These include ARTEP results, the records of Maintenance Management Centers, Equipment Readiness Reports, and IG inspection results. Whether a task contributes to critical combat systems has to be answered in terms of the judgment of experienced senior commanders.

Content Validity: This is a matter of task definition, task analysis, and congruence between task and scorable unit. Has the task been meaningfully defined? Have task performance requirements been comprehensively and rigorously analyzed? Does the scorable unit faithfully capture the key and essential elements of the task?

The determination of content validity is judgmental and is established in two ways:

(1) Review of the scorable unit against the task analysis on which it was based. Such a review is conducted both by the TDA and by ITEG.

(2) Review of the scorable unit by soldiers and their supervisors, who are asked to indicate whether the unit is a fair test of ability to perform the task.

Construct Validity: The term "construct-validity" is used here to convey the idea of validating scorable units against various types of evidence deriving from different sources. This would include the evidence bearing upon task criticality and content and concurrent validity described above, but it would include more than this. One type of evidence which seems especially relevant involves the analysis of pass/fail rates on scorable units in relation to external indices of task performance deficiencies. Suppose, for example, that a certain equipment maintenance task is tested on the SQT each year over a period of time, and that an initially high failure rate Army-wide shows substantial improvement. Suppose further that data are available on the incidence of equipment malfunction bearing upon this task. Are parallel trends evident in the two sets of data? That is, do indices of equipment malfunction improve as each related SQT performance improves? If so, it would be an indication of SQT validity. But this would do more than simply validate a scorable unit, for in the absence of a plausible competing explanation, such evidence would indicate that the SQT results do reflect real-world performance deficiencies and that the SQT is effective in focusing training efforts upon the correction of such deficiencies. This, of course, addresses SQT validity at the second of the three levels mentioned above, i.e., at the level of task scores and training management. In long range terms, this may be the level at which SQT validity is most critical. The evidence pertinent to validity at this level involves the analysis of SQT results in relation to performance indicators external to the SQT.

Now I wish to comment about validity with respect to the overall SQT score. As noted earlier, scorable unit scores are summed for each soldier to yield a total SQT score. In the distribution of SQT scores, two points are important: the qualification score and the verification score. The qualification score indicates that the soldier is at least minimally qualified in terms of technical competence for award of the next highest skill level. Except for waiver provisions, the soldier must achieve the qualification score in order to be eligible to compete for promotion.

The verification score indicates that the soldier is technically qualified to hold his present skill level. If the soldier fails to achieve the verification score, he is required to take the SQT again the following year. Failing again to achieve a verification score, the soldier may be reclassified, denied reenlistment, or subject to other adverse personnel actions.

Validity here concerns the soundness of these personnel management decisions. Were the soldiers promoted those who should have been promoted? Were those reclassified the ones who should have been reclassified? The validation of scorable units is clearly relevant to such questions.

If the scorable units have been shown to validly measure the ability of soldiers to perform critical tasks of their MOS, then this would seem to show that the SQT score has the type of job relevance which justifies its use in personnel management decisions.

The question being raised here is whether it is necessary or feasible to seek additional information bearing upon the soundness of personnel management decisions based upon SQT scores. Can these decisions be assessed in terms of future performance? Validation at this level presents major difficulties.

One difficulty which would have to be avoided has to do with the effects of personnel management decisions. That is, soldiers selected and those not selected for promotion, and soldiers selected for reclassification and those not selected differ from one another in the fact of these selections. Thus, we could wind up comparing soldiers who have been promoted with soldiers who have not been promoted, and soldiers who have been reclassified with soldiers who have not been reclassified, finally, proving only that nothing succeeds like success and finding out little about SQT validity.

This difficulty could be avoided by the use of waivers, i.e., through comparing soldiers who earned SQT qualification scores and were promoted with those who were promoted through waiver of the SQT qualification score. If a reasonably permissive policy on waivers is in effect during the implementation of the SQT program, such comparisons would be feasible. But they would have to be conducted in the face of other serious problems, including the extreme difficulty of finding useful and manageable indices of job performance to serve as validation criteria.

If soldiers who were promoted after receiving qualification scores are compared with those for whom the qualification score has been waived as a promotion requirement, on what basis are they to be compared? It could be in terms of supervisory and/or peer ratings, but these are fraught with problems of bias and unreliability. Conceivably, we could find more objective indices of performance, but I am not sure what they would be.

In view of the difficulty of finding satisfactory criteria of overall job performance, the validity of total SQT scores for personnel management perhaps can be best established through the validation of scorable units.

SYMPOSIUM:

Skill Qualification Tests: Measurement of the Occupational Competence of Soldiers

"MEDIA IN TESTING"

CPT ROBERT B. WILTSHIRE, USA
Army Individual Training Evaluation Group

MEDIA IN TESTING

The Skill Qualification Test (SQT) uses up to three performance oriented,criterion referenced methods for measuring an individual's job proficiency: a written component (WC), designed primarily to measure intellectual skills; a hands-on component (HOC) to measure motor skills; a performance certification component (PCC) to measure those skills which cannot be tested feasibly in either the WC or HOC due to resource constraints. Each of these three components requires a resource commitment, a standardized method of administration, and a standardized,effective presentation of job stimuli. Media can be used to reduce resource commitments, standardize the administration of the SQT, and present job stimuli more effectively than conventional methods.

Each of the three SQT components presents a unique set of problems. Perhaps the least impact on resources would occur with the PCC since the majority of resources required is found on the job. The presentation of job stimuli poses no major problem for the same reason. The problem of standardization, however, is probably greatest in the PCC. To reduce as much of the subjective scoring as possible, several strictures have been placed on PCC administration. All scorers are required to complete a scorer training program which includes a video tape presentation designed to illustrate proper and improper scoring methods. The tape stresses the importance of objective scoring. Additionally, each potential scorer must pass a test which judges ability to score examinees objectively. To further increase standard administration and reduce the number of test administrators, the use of an audio cassette to present examinee instructions is encouraged.

The hands on component might require a considerable investment in time, personnel, and equipment for some military occupational specialties (MOS). It is important to understand, however, that this investment should be viewed as an important contribution to a unit's training since the results of the SQT will not only be given to the individual, but also be compiled to produce a unit report from the company level up through the major command. Lower unit level reports will identify training deficiencies by task, while all unit reports will identify subordinate units with training problems.

As in the PCC, the HOC scorers must complete a scorer training program and the use of audio cassettes for standard presentation of instructions to examinees is again encouraged. The problem of presenting job stimuli is severely diminished since equipment and conditions found on the job are used in testing. The number of personnel used in scoring the HOC can be reduced through judicious use of media. In the task of

making a TA-312 field telephone operational, it was noted that a scorer at each of eight scoring stations watched the individual perform the task. Seven of those eight scorers could be eliminated by simply hooking all eight phones to a switchboard and having one scorer observe which station(s) made the correct connection. This "product" type scoring, or observing, the end result rather than the process which produced that result, can often reduce time and administrative personnel requirements.

Many tasks which are traditionally viewed as hands on such as trouble-shooting, are comprised of few motor and many intellectual skills. To illustrate, troubleshooting the NCR 500, a computer repairman's task, involves the extensive use of schematics as well as problem solving abilities. On the other hand, the only motor skills required are the hooking up of a multimeter and oscilloscope. Should the entire task be tested via the HOC or should the manual portion be tested in the HOC and the intellectual skills examined in the WC? Probably the latter since the intellectual skills lend themselves to testing in the WC and from a resource standpoint, scorers would not be required to stand during the HOC and wait while the examinee went through the troubleshooting thought process, but would only be required to determine if the examinee correctly connected the test equipment. The intellectual skills in this task could best be tested by giving the examinee the actual schematic used on the job and perhaps a cathode ray tube (CRT) terminal which enabled the examinee to sectionalize, localize,and isolate the fault using a light pen in place of the test equipment probes. The advantage of not having to induce faults in equipment which is delicate and/or must remain operational 24 hours a day can best be appreciated by the men who maintain and operate it. The number of tries with the light pen to isolate the fault can be recorded and used as a measure of the examinee's proficiency. The same result can be achieved without the electronic hardware using a paper based latent image technique which would also record the number of tries. Both approaches effectively test the intellectual skills associated with the troubleshooting. Media can also be used in the HOC to avoid disturbing operational equipment. A testing box designed by the U. S. Army Signal School can be adjusted to simulate equipment faults. Properly attaching test equipment to the box will allow the individual to determine the problem and at the same time, measure the examinee's ability to physically operate the equipment.

To standardize presentation of the WC, a package containing 35mm slides and an audio cassette will be sent to each Test Control Officer (TCO). Thus, a standard voice will be heard Army-wide accompanying a standard set of visuals. The program will be designed to accommodate questions from the examinees. This set of instructions should lower the number of errors made in filling out the mark sense answer forms the examinee uses to record responses for the SQT WC questions.

To present the examinee with "real world" cues in the WC, a description of an on the job situation is given to outline the requirements of the task. A series of questions is asked about the situation and from two to ten responses which parallel on the job alternatives are available to the examinee as answers to each question. Written situations, questions, and responses may not always be adequate, however. For example, one question submitted for review asked the examinee to determine the distance between the enemy and the individual's position prior to detonating a claymore mine. Only four distractors were given, each one a number of meters.

When carefully examined, the essence of the task was to accurately identify the position of the enemy force in relation to the kill zone of the claymore mine, a discrimination skill requiring depth perception. The task was not to identify a meter or memorize a number of meters. How then should the task be tested?

Going back to "What actually happens on the job?", revealed the most accurate test of the task was to have enemy troups charge a claymore position. Recognizing this was not feasible, an alternate method was devised using a video taped sequence showing an enemy force approaching the position. When the examinee felt that the unit was within the kill zone, he pressed a button stopping the videotape and recorded the number on the video tape recorder counter. Obviously, there is an effective range in which the detonation of the claymore mine would still produce an adequate kill. That range could be reflected in a series of numbers on the counter, say 100 to 110, which would be acceptable responses to the question. Using the above approach tests the task more accurately and effectively, and the need for the less desirable written description has been eliminated.

Another task, traditionally perceived as hands on, is the use of a speed trap. This task performed by Military Police (MP), uses mirror boxes. It is, however, no longer commonly employed since more sophisticated approaches are available. The task serves, however, as a good example of media use in testing.

A car approaching the military police position will be reflected in one of two mirror boxes. When the car is reflected in the first mirror box the MP presses a stop watch and waits for the car's reflection in the second mirror box. When the MP sees that reflection, he depresses the stop watch, notes the time elapsed, and determines if the car was speeding. Running cars by MP's standing by a speed trap is an expensive proposition. The Military Police School, however, developed a video tape which effects the same thing. An examinee can now sit in a carrel, watch a video tape of cars passing a speed trap, hold the stop watch, observe the cars' reflections, and determine if any car

was speeding. Once again, the real world cues have been effectively presented and the examinee's actions closely parallel those on the job. Why video tape? It is a flexible medium which provides motion, can be edited much more easily than film, and can be completely erased and reused.

The use of media for testing is by no means confined to videotape. Pictures, slides, film, realia, and other media might all be used in the testing arena. The use of illustrations in an SQT WC for example, is highly encouraged for two reasons. First, it presents the examinee with the visual job stimulus and second, it significantly reduces the requirement for an individual to be a good reader or visualizer. The person who does not have adequate reading ability but who can perform the job should not be penalized. Illustrations help overcome that inequity. Correct and incorrect sight pictures for the M16 rifle, correct and incorrect alignment for air defense radar, correct and incorrect safety techniques, pictures, and diagrams are all examples of effective illustration use.

Closely akin to illustrations are excerpts from the training literature that an individual has on the job. A chart in a technical or field manual which must be used in the performance of a task can be exerpted for the WC to produce a performance question. The question will require the examinee to refer to the chart just as he does on the job, in order to answer the question and demonstrate job proficiency. An individual should not be required to memorize any portion of a manual unless the job calls for memorization. The use of excerpts and addenda to the SQT must be tempered by practical limitations, however. Obviously, entire manuals cannot be appended to the WC test booklet. But manuals can and should be used in the HOC when necessary for correct operation of equipment.

A concept called Group Paced Testing is being used in several SQT. It will require simultaneous participation by examinees on time constrained tasks such as aircraft or tank recognition. Under this concept, a series of 35mm slides will be presented to the test group, each slide for a specified number of seconds. That slide will be followed by the next slide, and the next until all slides are shown. Since all examinees will be limited to the same amount of time for the series of slides or scorable unit, it is more efficient to administer the presentation to a group, than to provide each examinee with a set of slides.

In addition to the traditional individual and group paced testing format, a third possibility, self-testing, exists. In this instance equipment presently in the field or programmed for future distribution such as the Sony Rover (single camera, videotape recorder), and an audio recorder with cassettes, can

be configured in a room to produce a self-testing facility. The following example demonstrates how the facility would operate. The task is to use the lensatic compass for day navigation. The examinee is given a set of instructions prior to entering the testing facility. The instructions tell the individual to turn on the tape recorder and follow the instructions given. Those instructions tell the examinee to start the single camera videotape operation, to stand on a premarked position on the floor in front of the camera, to demonstrate the correct method of sighting the compass, to look at premarked points around the room, site on those points as they are given from the tape recorder, give the readings as indicated by the compass to another tape recorder, the video tape recorder, or write them down. The individual then turns off the equipment and resets it for the next examinee. Why the bother? Well, first of all, it gives the individual a chance to test his skills at his convenience, and secondly, it provides a permanent record which would be available for him and an evaluator to review. Needed corrections could be graphically displayed and the test would serve as a valuable learning tool. The flexible scheduling would allow an individual to be tested at his or a commander/supervisor's convenience thereby using available time more efficiently.

The above example alludes to testing as more than just an external evaluation tool. A pre-test is used to determine if an individual has the requisite background for a course, or has a background which obviates the need to take a portion or all of the course. A post test is used to ascertain if the individual has mastered the skills taught in the course. Imbedded testing is used in a training program to measure an individual's progress and provide him feedback on how well he is mastering the objectives. Media can be used effectively in all three instances just as it has in the examples given for the SQT, an external evaluation method. Imbedded testing more than the other varieties, however, serves a remedial function. Using again the example of the lensatic compass, a self-instructional package which guides the learner through day navigation can be accompanied by a videotape which serves as a review and a test of the elements just learned. In this instance the individual would observe a master of the task sighting a compass and taking readings. The videotape asks the individual to duplicate the master's actions. If the individual obtains the correct reading or sight picture as demonstrated on the videotape, he passes the test. If not, he continues practicing. Imbedded testing could also include the flashcard technique using 35mm slides or traditional flash cards, to test the learner's ability to discriminate the compass parts used in sighting. The application of effective media use in testing is limited only by the developer's imagination.

Whatever medium is chosen for testing, the choice should be based on the best medium or media to test the given skill, and practical limitations. The best medium may be videotape but equipment nonavailability, cost, or any other variable might preclude its use and dictate the adoption of the next best medium. The most important consideration is that the individual reach the required standard of proficiency as quickly as possible with maximum retention. The effective use of media will contribute significantly to that end.

A written test is traditionally paper-based and requires a monitor to distribute and collect test booklets and answer sheets, package and forward the answer sheets to a central scoring facility, and destroy and account for the test booklets. Use of state of the art hardware and software could eliminate a paper based test. The possibility of having a test delivered to an individual in a carrel replete with a CRT terminal, a video tape player, an audio cassette player, a 35mm projector, and auxiliary equipment is quite real. The computer terminal could be linked to a centralized library via a cable network. The library would contain training materials to include tests, that could be transmitted to any CRT terminal on a demand basis. Responses to test questions would be immediately graded and incorrect answers would trigger a display showing the correct answer and a reference where that answer could be found. Additional equipment could produce hard copy test results and prescriptions for further training. This feedback could be used by the individual and anyone having responsibility for the individual's training and education. To complete the system, the individual's progress and test results could be stored and forwarded to the next unit of assignment via the personnel computer system so that the gaining commander/supervisor would have specific knowledge about the individual's capability to perform his job and his training and education could proceed uninterrupted. Since retention decay rates vary from task to task and from element to element within a task, retraining and periodic retesting of the individual would be required. A computer which keeps track of an individual's training program and progress could schedule retraining/testing according to empirically established decay rates. A computer link to the personnel system would provide "for record" test results and educational progress (i.e., high school completion, college credit) which could be used in consideration for award of promotion.

Obvious advantages to the centralized library and individual carrel concept include immediate feedback and training prescriptions. The economic advantage of significantly reducing a paper based testing and training system is great. The cost of paper for test booklets and answer sheets would be eliminated. So too would the expense of mailing them be gone. The costs associated with printing, handling, and pro-

cessing would be eliminated. The error rate of filling out header information on a mark sense test answer sheet would be greatly reduced since the majority of personal information about the individual would already be on file and there would be no need to enter that information on each mark sense form. Another advantage is the elimination of a test monitor to distribute test packages, read instructions, assemble tests and answer sheets, and forward them to a centralized facility. This could all be accomplished by a combination of media.

What about the cost of such a system? By carefully incorporating in the system existing hardware and hardware programmed for distribution Armywide, such as that associated with the Training Extension Course program, the Army could eliminate a portion of the front end cost. The remainder of the system could be phased in over a period of years as the existing paper based system was phased out. Monies used to support the present system could be transferred to defray the one time front end costs of the new system. Amortization of one time equipment costs over x number of years, might well prove the new system less expensive than the present one. In any event, thorough costing of each system and a long range benefit analysis must be completed before any final decision is made.

This discussion may sound Buck Rogerish to some, but it can be done. In fact, it has been done via satellite in 1974, between Denver, Colorado and a V. A. hospital in Dublin, Georgia.

The use of media has a definite place in the world of testing. In the limited time allotted in this conference, only a brief glimpse of the use of media in the SQT and an even briefer glimpse of a proposed system to increase training effectiveness and management, was possible. The examples given in the presentation illustrate how media can be used to reduce resource commitments, standardize the administration of the SQT, and present job stimuli more effectively than conventional methods. The tip of the ice berg has been barely touched but through creative thinking, imaginative research, and a sharing of ideas, testing which accurately measures an individual's ability to perform his job can become an effective and efficient realization.

CPT ROBERT B. WILTSHIRE

INTRODUCTORY REMARKS FOR THE PANEL ON

THE INSTITUTIONALIZATION OF CHANGE IN MILITARY TRAINING

Worth Scanland, PhD.,
Staff
Chief of Naval Education and Training

It is well known, or at least assumed, axiom in the educational establishment that it requires approximately fifty years for the institutionalization of a significant change in public education. That length of time may have been reduced somewhat in the last few years, as the American people have become more accustomed to technological changes, but probably not to any great extent. Now, we might easily deceive ourselves into the belief that because the military establishment is basically an authoritarian institution, it would be relatively easy to effect change...after all, we really need only to have the commander issue the appropriate order and whatever change is necessary will occur. Unfortunately, my fellow trainers, this is not the case, as some of us who have attempted the introduction of innovation have painfully observed. Because the practice of instruction, be it in education or training (if there is a difference), is a social science, we have discovered that the introduction of significant change has had social impacts. Social impacts implies, among other things, changing roles and self perceptions, certainly two of the most powerful of human characteristics. Therefore, the institutionalization of change in military training is just as traumatic as is the institutionalization of change in less authoritarian institutions, and because these are times of rapidly changing technologies in the science of instruction, it becomes extremely important for educators and trainers to study ways and means for effecting these changes. So it is my pleasure and privilege to introduce a panel of authorities in the subject of change, and we will listen to four perspectives which we believe to be relevant and important to the topic. These discussions follow.

THE ORGANIZATIONAL AND POLITICAL PROBLEMS OF IMPLEMENTING CHANGE IN GOVERNMENTAL ORGANIZATIONS

Lawrence J. Korb
Naval War College

INTRODUCTION

The problem of implementing change is one of the most difficult faced by a leader or a decision maker. There is a wide literature on the subject of organizational change. This paper will not attempt to add any new insight into the psychological and attitudinal theories of change. What it will try to do is to point out some perspectives that are often overlooked by those concerned with change in government organizations.

These perspectives are those of a political scientist and are based upon the pioneering work of Harvard political scientist G. Allison. In his 1969 article in the American Political Science Review and his book on the Cuban Missile Crisis, The Essence of Decision, Allison pointed out that there are three models which can be used to explain organizational behaviors or outputs: the Rational, Organizational and Bureaucratic Models or Models I, II, and III.

Although these models are applicable to the entire decision making process and not only to the problem of change, I would argue that they can be most helpful in understanding the problem of implementing change. In this paper I propose to demonstrate this by focusing on two current situations in which policymakers are trying to bring about change in military training: Marine Corps Recruit Training and Navy Helicopter Training.

Model I - the Rational Model

Model I or the Rational Model assumes rational or purposeful behavior on the part of all of the individuals involved in the policy making process. Rational behavior is more than merely just intelligent behavior. It implies behavior motivated by a conscious calculation of advantages. This calculation in turn is based on an explicit and internally consistent value system.

Most change makers or change agents implicitly or explicitly adhere to this model. They assume that decisions are made by rational agents who will choose among a set of alternatives, fully aware of the consequences of each. Those advocating change also assume further that the affected individuals and groups will examine the costs and benefits of each alternative, will explore the various options in light of the

pressures in the environment, and will select the value maximizing course. Thus, if someone were interested in getting the Marine Corps to implement changes in their recruit training procedures or the Navy to accept a joint helicopter training program under the control of another service, he would assume that it would be sufficient to demonstrate to the Marines and the Navy the rationality of those courses of action. For example, these changes would enable the Marines to solve their recruiting shortfall without lowering their standards and the Navy to train adequately the same number of pilots for a given cost. While this strategy might work in some cases, the chances are that it would not be successful because the rational approach does not fully account for the organizational and political realities of governmental organizations.

Model II - the Organizational Model

The organizational model in contrast to its rational couterpart emphasizes the fact that individuals are also members of groups or organizations, e.g., Navy, Marine Corps, Naval Aviation, and as such, have been socialized to some degree into accepting specific organizational views and behavior patterns which unconsciously condition the way in which these individuals respond to stimuli. The impact of organizational norms upon the response of individual members has three interrelated aspects.

First organizational issues become the criteria for defining acceptable behavior by group members. Thus, in evaluating courses of action, the individual looks primarily at what is good for maintaining the essence of the organization rather than the benefits for the society at large. Air Force Officers almost unanimously oppose the proposal to eliminate the manned bomber inspite of evidence which proves that missiles are more cost effective in carrying out the strategic mission. Their opposition stems from the fact that the manned bomber is closely associated with the essence or specialty of the Air Force. Similarly, many Marines oppose relaxation of their recruit training procedures because this training is part of the essence of the Corps. Without it, the Marine Corps recruit could become indistinguishable from the Army recruit.

Second, governmental organizations develop Standard Operating Procedures (SOPs) for dealing with routine issues and scenarios to deal with various contingencies. After a period of times, these procedures become so standard or ingrained that they are not susceptible to quick or easy change. In some cases, members can come to associate these procedures with organizational essence.

Third, when confronted with situations to which either the SOPs or the standard scenarios are not applicable, the organization engages in a local search. That is, they do not really consider all of the available alternatives but only those which are similar to or marginally different from the standard procedures already developed and which do not do violence to the organizational essence.

Thus, when tasked to change the pattern of recruit training, the Marines would probably not propose any dramatic alternatives or would suggest only those which would bring about marginal changes in existing procedures. If such changes were proposed by an outside group, many members of the Marine Corps would resist those changes for fear they would do violence to their essence or specialty. From the perspective of the Marines, relaxation of recruit training would make them similar to the Army and could lead to their absorption by the Army, as nearly happened in 1947 and has again been proposed in 1976. Similarly, the Navy would view joint helicopter training as a threat to Naval Aviation which they regard as part of their organizatonal essence and a potential first step toward the loss of their air arm. If changes were imposed by outside agencies, i.e., the Secretary of Defense, there would be delays in implementation and a tendency to make cosmetic rather than real changes.

Model III - the Political Melieu

By definition governmental organizations are involved in a political melieu. As James Forrestal, the first Secretary of Defense so aptly phrased it, "...you can no more divorce government from politics than you can separate sex from creation." Therefore, the behavior of government actors is never far removed from political considerations. To understand the impact of politics upon behavior in a given situation, four questions may be asked: Who is involved in the situation; i.e. who are the players? What determines each player's stand? What determines each player's influence? How does the game combine players' stands, influence, and motive to yield decisions and actions, i.e., behavior?

The question of who plays can be answered by identifying the individuals in positions and jobs in the government that produce action on the questions involved. Position defines both what the individual player can and must do. For example, does a change in Marine recruit training involve the Commandant or can it be handled on a local or base level? Does the change require approval of the Secretary of the Navy or can it be handled internally by the Marine Corps? Is this a decision which is likely to attract the attention of the Secretary of Defense, the White House, the Joint Chiefs of Staff, or the Congress? Similar questions could be raised about the level at which the Navy decision on joint helicopter

training will be made. The answers to the question of who plays will shed much insight into how much change will be implemented.

Each player's stand will be determined by a number of considerations. The parochial priorities of his group will undoubtedly color his perception of the issue. The stakes involved will affect the position he supports. The stakes normally include national security, organizational, domestic, and personal interests. Finally, deadlines will affect his stand. Within government, deadlines are brought about by such events as budgetary cycles and major speeches. Deadlines imposed by such happenings not only force decisions but also determine the forums or channel in which the issue is raised.

Thus, model III, or the political model, cautions that the stand of the Commandant of the Marine Corps on changing recruit training will be influenced by such things as concern for the organizational health of the Corps, his own background, when, where, and with whom the issue is raised, and the deadline by which a decision is needed. Neither the Secretary of the Navy, the President, nor Congress is likely to see the same stakes and focus of the issue as the Commandant. Similarly, the CNO, the Chief of Staff of the Army, and Congress likely will view the question of joint helicopter training differently. Changing Marine Corps training methods can be done within the Executive Branch; creating joint helicopter training requires Congressional approval.

The influence or power of each individual in any particular area is an elusive blend of bargaining advantages, skill and will in using bargaining advantage, and other players' perception of the first two ingredients. The source of bargaining advantages includes such things as formal authority and responsibility, control over resources necessary to carry out action; expertise and control over information needed to define the problem; control over information that enables leaders to determine whether and in what form decisions are being implemented; the ability to affect other players' objectives in another area; personal persuasiveness; and access to and persuasiveness with players who have bargaining advantages.

Model III urges the change maker to discover who are the players, what are their stands, and what is their influence; and then to construct a strategy to bring about a desired solution. The political model cautions that those who ignore these considerations are unlikely to implement the desired change. The Marines are not likely to change recruit training by themselves. It will require pressure from above, possibly from the Secretary of the Navy or the President who have some bargaining advantages or influence over the policies of the Corps. Likewise, those who wish to convince the Navy to accept

joint helicopter training would be wise to construct a strategy to overcome the opposition of the Navy and the influence of certain powerful congressmen.

CONCLUSION

Individuals in governmental organizations who are involved in the decision making process do not act irrationally. However, their behavior is not motivated solely on the basis of economic rationality. Organizational and political factors condition their responses to change proposals. This suggests that the change maker must be aware of all these models or that there is an implicit fourth model which combines the perspectives of Models I, II, and III.

THE PRAGMATIC ASPECTS

OF INSTITUTIONALIZING CHANGE

FRANKLIN F. SANDS, PH.D.
HEAD, APPLICATIONS OFFICE
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA

PAPER PRESENTED AT THE 18TH ANNUAL CONFERENCE OF THE
MILITARY TESTING ASSOCIATION
GULF SHORES, ALABAMA - 18-22 OCTOBER 1976

THE INSTITUTIONALIZATION OF INNOVATION IN MILITARY TRAINING

The Pragmatic Aspects of Institutionalizing Change

ABSTRACT

Most large organizational entities and indeed the military services, continually experience simultaneous and sometimes conflicting pressures for both stability and change. The increased pace at which change (scientific, technological, societal, etc.) is occurring today has directed the attention of managers toward the need for creating an organizational climate to successfully accommodate the implementation of the changes that have become necessary for the growth and survival of their organizations. Institutionalizing new technology within military personnel management, education and training agencies involves risks that tend to signal some practical implications which can be as important as the implications of the new technology itself. This paper addresses some of the pragmatic issues raised when planning for the effective utilization of human resources research end products.

Indulge me if you will this morning, while I make several sweeping and perhaps fragmented comments. They speak to my perception of some of the more pragmatic issues that deserve attention when attempting to institutionalize new technology. They speak to the conceptual label which institutionalization represents. They speak to a need for breaking-out and coming to grips with the more tangible issues subsumed under this label. Within the context of a problem-solving paradigm, they also speak to some basic steps which should be included when planning a coherent strategy for achieving a desired operational state of affairs.

Identification and Diagnosis of the Problem

Effective change strategies tend to be systematic and begin with the recognition that a particular program, a functional organ or a man-weapon system interactive mix is malfunctioning. This may appear self-evident to some, however, many problem-solving efforts fail because responsible change agents have difficulty in ascertaining that a given set of circumstances represent an actual problem; they tend not to be sure just what problems should be tackled; or even who the potential beneficiaries will be, if the problem is solved.

Planning decisions are, therefore, very important since they are generally designed to undergird objectives toward which change actions are directed. They specify major changes that are needed in a program. The need for planning decisions usually comes via the recognition that disagreement exists between what a particular program was intended to be and what it actually is; or recognition of disagreement between what a

particular program can become and what it is likely to become. Planning decisions in either case could be, or should be, made to change or not to change either of the intended state of affairs or the actual state of affairs. Therefore, any such decision to introduce change would result in the establishment of program planning objective, structural objectives, and implementation objectives which query: whether current program goals should be changed?; whether the present mission is accurate and should be sustained, or should be changed?; what are the top priority needs that are or are not being served by the current program?; what is the nature of, or what are the characteristics of the problems to be solved in accommodating the top priority needs? The fundamental importance of accurate problem identifications and diagnosis, which result in planning decisions, is that of providing crucial information for the formulation of new operational objectives or modifications of existing objectives.

Diagnosing the problem, searching for alternative solutions, testing of solutions, and implementing those alternatives selected with subsequent evaluation and feedback, describe some of the processes through which a systems analysis must proceed. However, the constraints and barriers which surround defense systems must be carefully understood before a potentially successful change strategy can be formulated.

(RE) Searching for Solutions and Relevant Knowledge

Having successfully diagnosed the problem situation, and gained substantial clarity of the diagnostic findings, the next step would involve a search for relevant knowledge and solution alternatives. Such a step

involves the introduction of three practical requirements: 1) becoming aware of potential solutions through systematic review of appropriate information sources; 2) choosing from an array of promising alternatives the one solution which best fits the problem at hand; or 3) developing a valid and viable operational R&D requirement with an objective of advancing or extending the relevant knowledge base, which in turn may provide a basis for innovation.

Let us look at some of the practical issues pertaining to research and development.

The support which the research, development, testing and evaluation enterprise now enjoys is provided on the explicit assumption that research and development can significantly impact on the operational readiness of military forces. This assumption of operational utility of RDT&E has, over the years, undergirded and sustained the growth in research and development support. Now, however, the assumption is coming under increasing challenge by policymakers at the Department of Defense and in the Congress. Considerable interest and emphasis is being placed on making the research organization and thereby, the researcher "accountable" and making the research products "useful." Increasing numbers of critics are asking for hard evidence of what research has done for the Navy recently. The attitude held by many would suggest that too much R&D money has gone into poorly conceived projects with too few results being vigorously assessed for purposes of implementation. Like others, we at NPRDC have experienced this change in attitude and recognize several situations which tend to have a fostering effect.

Important progress, which is actually occurring as a result of RDT&E programs, has not been adequately communicated to the policymakers who now are responding to budget pressures.

A second situation may also be that the total dollar investment in human resources RDT&E is simply too small, restricting any provisions for adequate diffusion and utilization.

Still another situation concerns the prevalence of general application problems in the RDT&E organization such as:

(1) The tendancy to prolong research programs and keep them in research versus implementation.

(2) Turnover of consumer personnel (research requirement is specified by one set of persons; however, research results are delivered to a new set of persons).

(3) Poor communication techniques employed before, during and after research projects.

(4) Insufficient planning to anticipate cost (to consumer) of implementing research results.

(5) Non-implementable, inconclusive, or unsatisfactory research (as perceived by the consumer).

(6) Consumer located at too low an organizational level to implement results particularly those requiring funds expenditure.

(7) Many of our consumers need help and guidance from the research establishment to implement findings.

We also recognize a need for having a viable in-house RDT&E follow-on system.

A system having the sensitivity to descriminate between research efforts designed to increase and refine our technology base, in the interest of future development, and research efforts designed for demonstrable application.

A fifth and similarly perplexing situation concerns research which is "useful" but not "used". We recognize that high-quality RDT&E capability and high-quality RDT&E end products do not guarantee utilization.

For these reasons we have initiated an application program as a function of the total RDT&E effort, designed to facilitate the introduction of end products into operational use; to develop and employ an application model consisting of communication and coordination networks to determine the utilization status of research findings and products, and to develop new methods and techniques for research dissemination and utilization through comprehensive study. These efforts should improve the researcher/sponsor/operator relationship and represent the beginning of an improved coordination process extending from initial problem determination through product implementation and feedback.

There will, however, remain some rather practical problems in implementing R&D products such as:

(1) The more "hard" or physical a product, the more readily it is accepted. This results because the decision maker can "see" what it will do for him. Physical developments usually pose no threat to the manager as a decision maker, except for some intelligent, automatic system.

(2) The more "soft" the product--i.e., the more it deals with people or "human factors"-- the more difficult it is to implement. This means that the decision maker cannot "see" what it will do for him and, therefore, he is highly ambivalent about it, at best. It may also be perceived as a threat to his managerial prerogatives and perhaps denigrating to his abilities as a manager. Also, he may consider himself to be a better authority in the area of the product than the scientist.

(3) The more new technology involves a change in the way of doing things, the more difficult it is to implement--the innovation problem. The prevailing norm is, "Don't rock the ship." Obviously, security is upset by major changes. Even hardware changes are difficult to implement if they involve a major change in organization, personnel, and procedures. Small improvements in an existing system are easier to implement than a radical substitution of systems. Major changes also require the operation of a parallel system while the changeover takes place, along with the attendant increase in expenses and confusion.

(4) Implementation of R&D products tend to be easier at the lower levels of the organization. Again, this would seem to be rooted in the fact that the incumbent can see the improvement it would make in his way of doing things. There is less likelihood of the change being a threat to his managerial prerogatives. This suggests a strategy of sneaking in innovations by the back door so that by the time they reach the attention of management, there is a fait accompli and implementation is essentially completed.

(5) Implementation, when it occurs, tends to take place rather late in the R&D cycle. A consensus or climate of acceptance must prevail in the managerial, decision-maker community before major R&D products are implemented. A corollary of this assertion is that implementation of R&D products, when it occurs, tends to occur in a rush--a bandwagon effect. For example, when the information revolution came about, everybody wanted to "computerize" it; a similar reception greeted such areas as programmed learning and microminiaturization.

(6) Products of R&D that involve immediate, short-term solution to current problems are more readily implemented. Most payoffs to management are based on current performance. In some cases, there is not much benefit in instigating reforms or programs that will bear fruit in another person's tenure. This is especially true in the military where superior and subordinate transfers occur routinely and at short intervals. Short-term solutions capitalize on currently available, "off the shelf" R&D products that have high visibility and general acceptance in the decision maker's community. Long-term projects that compete directly against an available short-term product being implemented will offer the most.

Solution decisions to mobilize toward change tend to be more directly influenced by the D in the R&D cycle. For it is the development activity, and not research, which is crucial to the process of change. For while research may make change possible, solution decisions and adoption strategies are driven by that level of development which produces the innovation to be adopted.

Mobilizing for Change

Mobilizing for change would involve activities such as:

Dissemination, in order to create wide spread awareness among potential users of the new technology.

Demonstration, in order to afford an opportunity for assessing the new technology's credibility, relevance, relative advantage over other products, ease in understanding and installing, and compatibility with established values, norms, procedures, and facilities.

Training, in order to insure that the user has all of the essential skills attendant to the use of the new technology.

Installation, in order to operationalize the use of the new technology by fitting its characteristics to the adopting organizations' environment.

and lastly, Marketing, in order to promote installation of the new technology, based on a feasibility and preferability analysis. These analyses should illustrate whether the new technology will work and at what cost?; and how well the new technology works and at what comparative cost?

Some other practical concerns which impact on installation decisions and the formulation of stabilizing strategies would include:

Economic aspects - how much will it cost and can the particular command affort it?

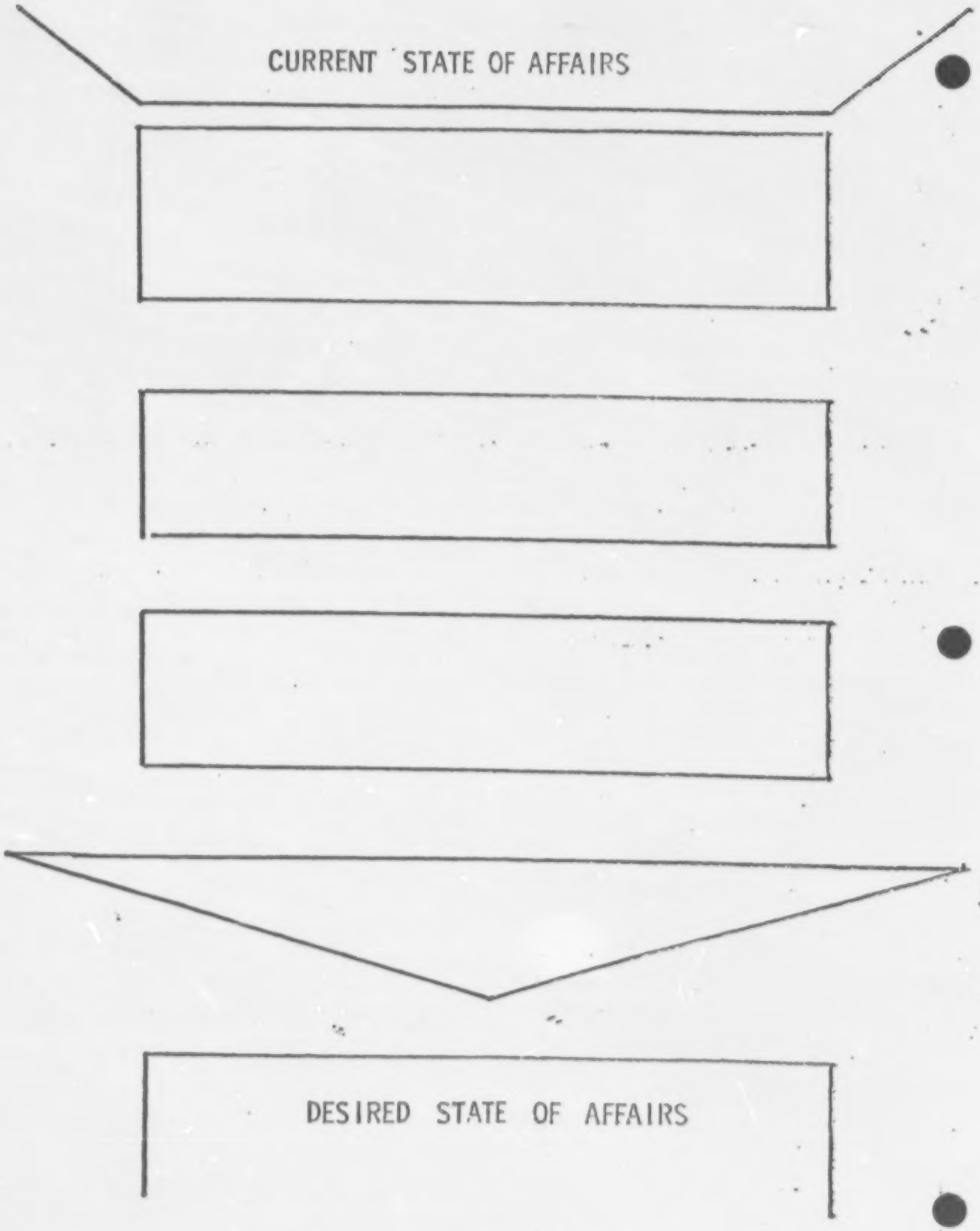
Technical aspects - is the talent available, and are the required facilities accessible to fully implement the new technology in order to derive the desired outcome?

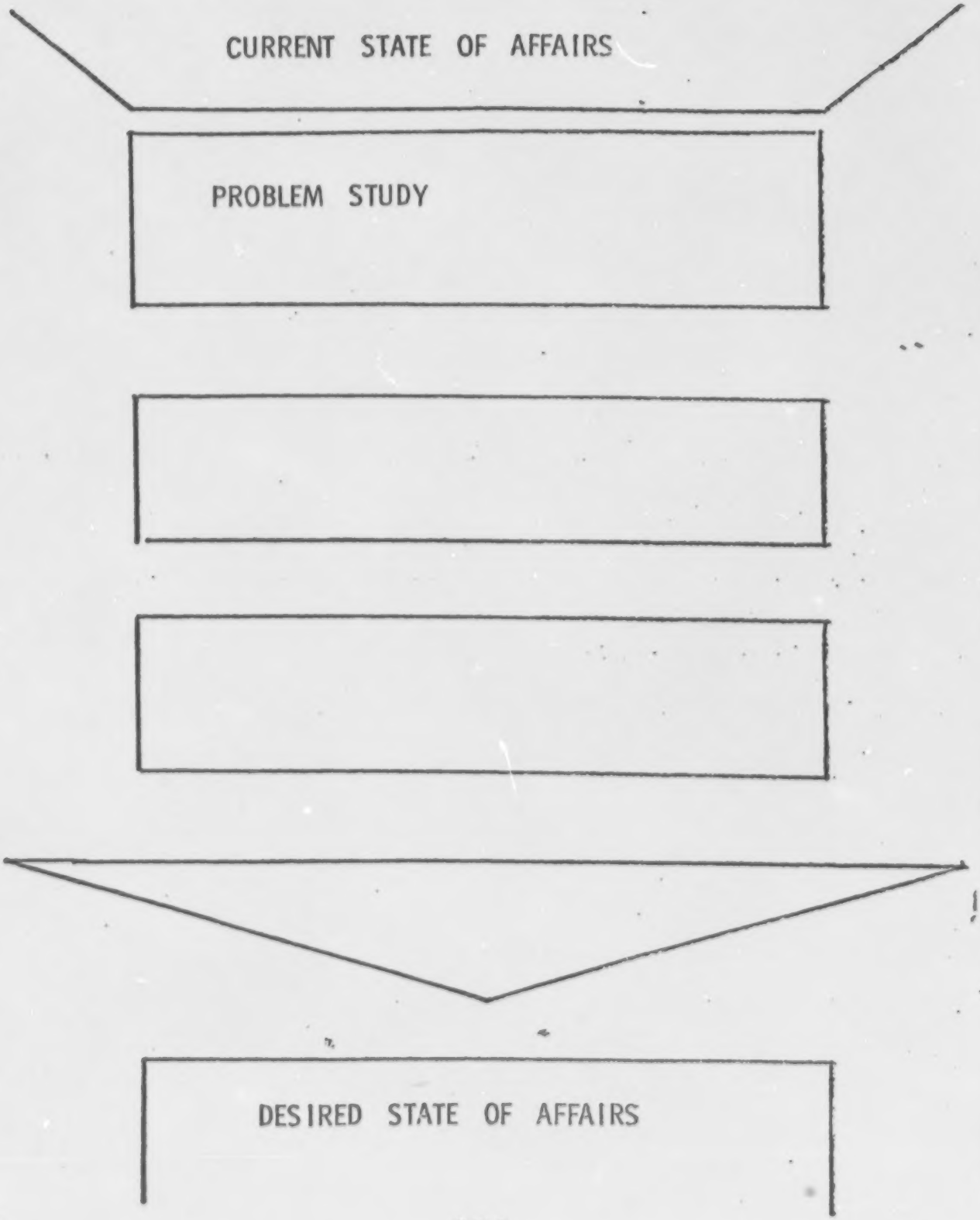
Political aspects - is there general acceptance of the new technology outside the immediately affected boundary? Is it politically prudent to endorse the introduction of the new technology? Is the new technology humane?

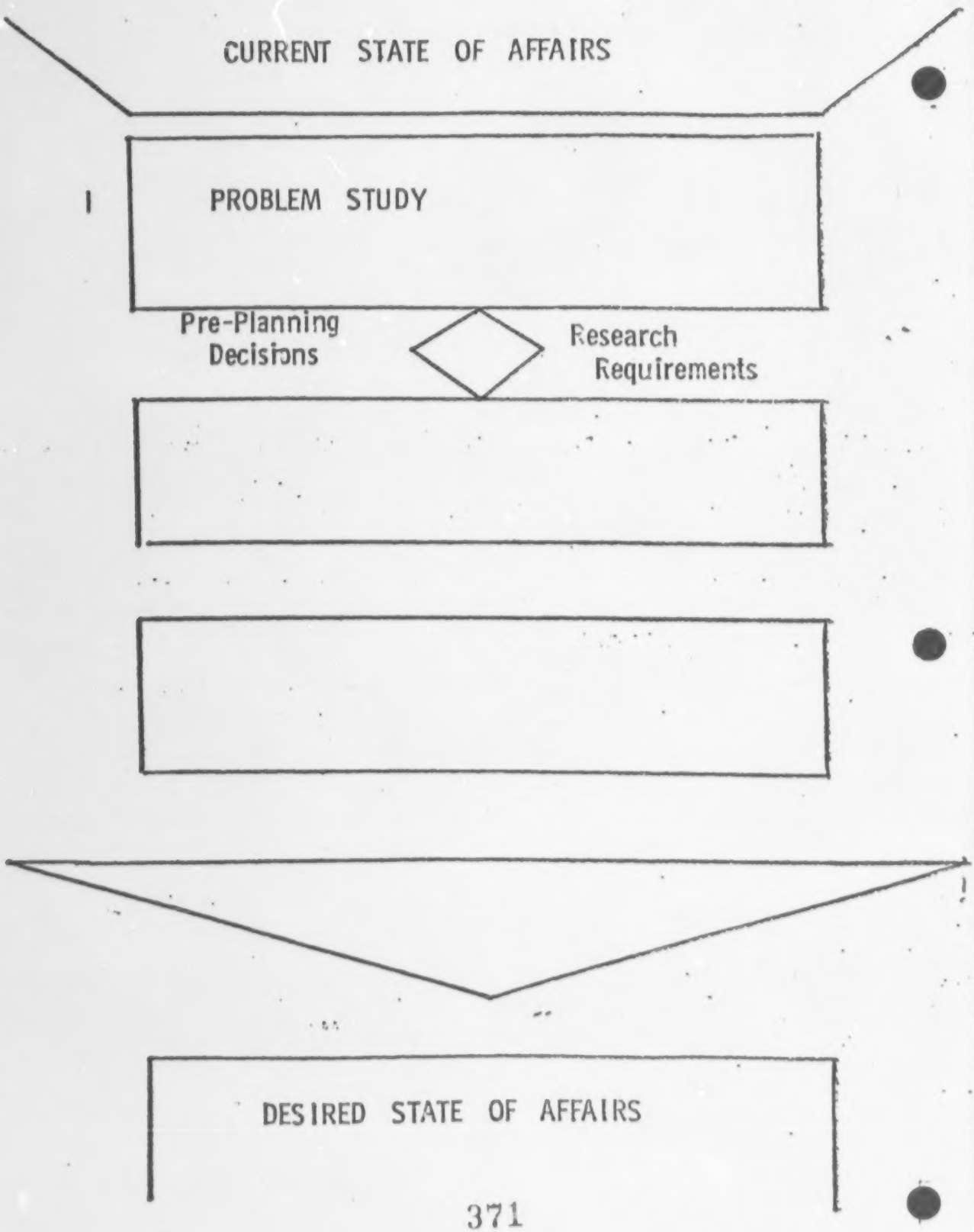
Institutionalizing Change

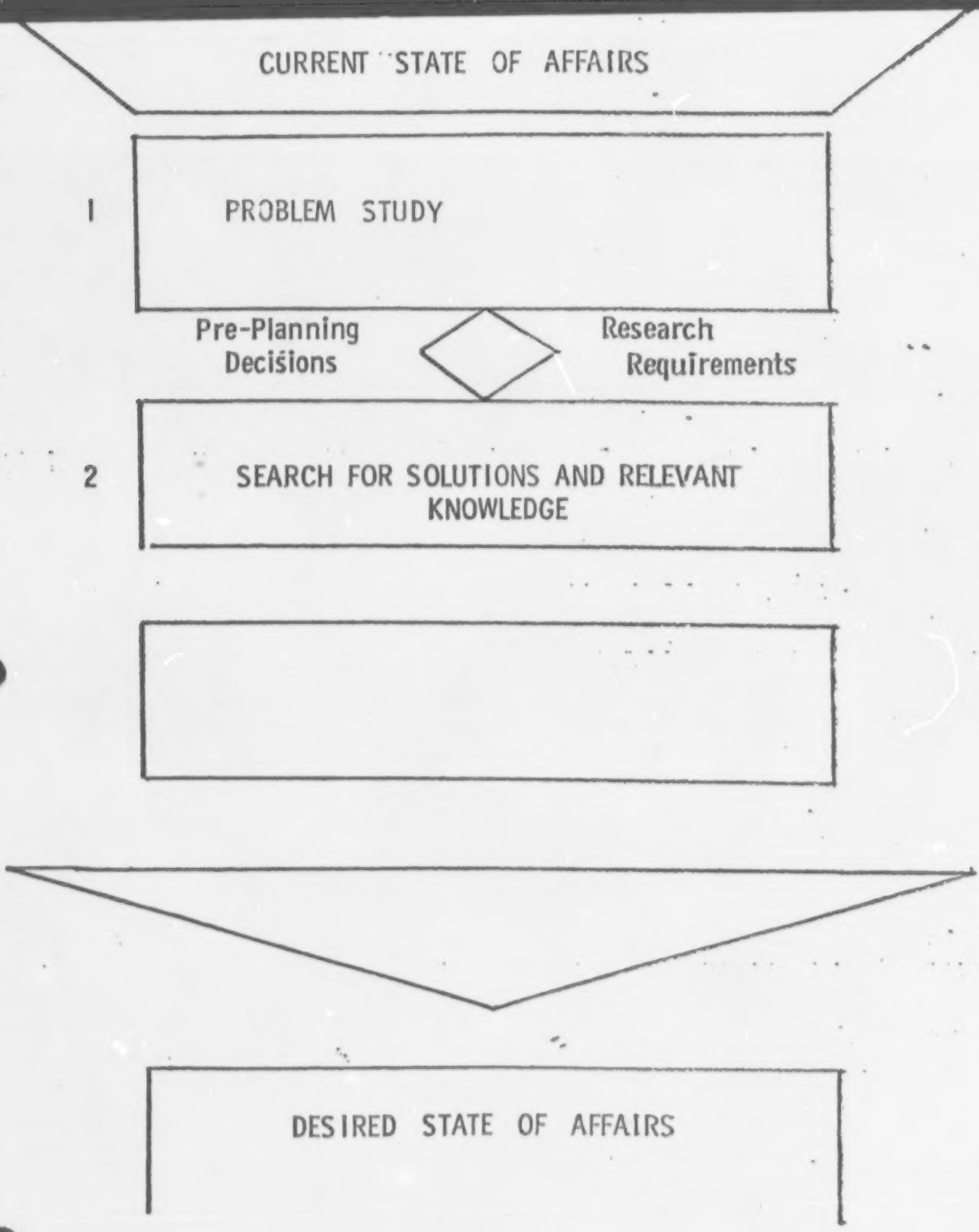
The adoption of new technology, or the change process will have completed a cycle when the new technology is assimilated as an integral, accepted, well established part of the operation. Operational stability, operational continuity, a high level of organizational support for the new technology and the degree to which the new technology is valued are important aspects in assessing this conceptual state of change - that is, when new technologies are known to be institutionalized.

We should...indeed we must come to grips with the pragmatic aspects involved in effectively managing the institutionalization of technological innovation, if we are to derive maximum and timely benefit from those changes that are necessary for operational growth.


CURRENT STATE OF AFFAIRS
DESIRED STATE OF AFFAIRS


CURRENT STATE OF AFFAIRS
PROBLEM STUDY
DESIRED STATE OF AFFAIRS


CURRENT STATE OF AFFAIRS
PROBLEM STUDY
Pre-Planning Decisions
Research Requirements
DESIRED STATE OF AFFAIRS


CURRENT STATE OF AFFAIRS
1
PROBLEM STUDY
Pre-Planning Decisions
Research Requirements
2
SEARCH FOR SOLUTIONS AND RELEVANT KNOWLEDGE
DESIRED STATE OF AFFAIRS

PRODUCT IMPLEMENTATION OBSTACLES

- INADEQUATE COMMUNICATION OF IMPORTANT PROGRESS
- TOTAL DOLLAR INVESTMENT IS TOO SMALL
- GENERAL RESEARCH APPLICATION PROBLEMS
- INADEQUATE IN-HOUSE RDT&E FOLLOW-ON
- RESEARCH JUDGED USEFUL, BUT NOT USED

APPLICATIONS PROBLEMS

1. TENDENCY TO PROLONG RESEARCH PROGRAMS AND KEEP THEM IN RESEARCH (VS IMPLEMENTATION) PHASE.

2. TURNOVER OF CONSUMER PERSONNEL. RESEARCH REQUIREMENT IS SPECIFIED BY ONE SET OF PERSONS, RESEARCH RESULTS DELIVERED TO A NEW SET OF PERSONS.

3. POOR COMMUNICATIONS TECHNIQUES BY RESEARCHERS AND RESEARCH ORGANIZATION BEFORE, DURING AND AFTER RESEARCH PROJECT.

4. INSUFFICIENT PLANNING TO ANTICIPATE COST (TO CONSUMER) OF IMPLEMENTING RESEARCH RESULTS.

5. NON-IMPLEMENTABLE, INCONCLUSIVE, OR UNSATISFACTORY RESEARCH (AS PERCEIVED BY CONSUMER).

6. CONSUMER AT TOO LOW AN ORGANIZATIONAL LEVEL TO IMPLEMENT RESULTS, PARTICULARLY THOSE REQUIRING FUNDS EXPENDITURE.

7. CONSUMER NEEDS HELP/GUIDANCE FROM RESEARCH ESTABLISHMENT TO IMPLEMENT FINDINGS.

PRODUCT IMPLEMENTATION OBSTACLES

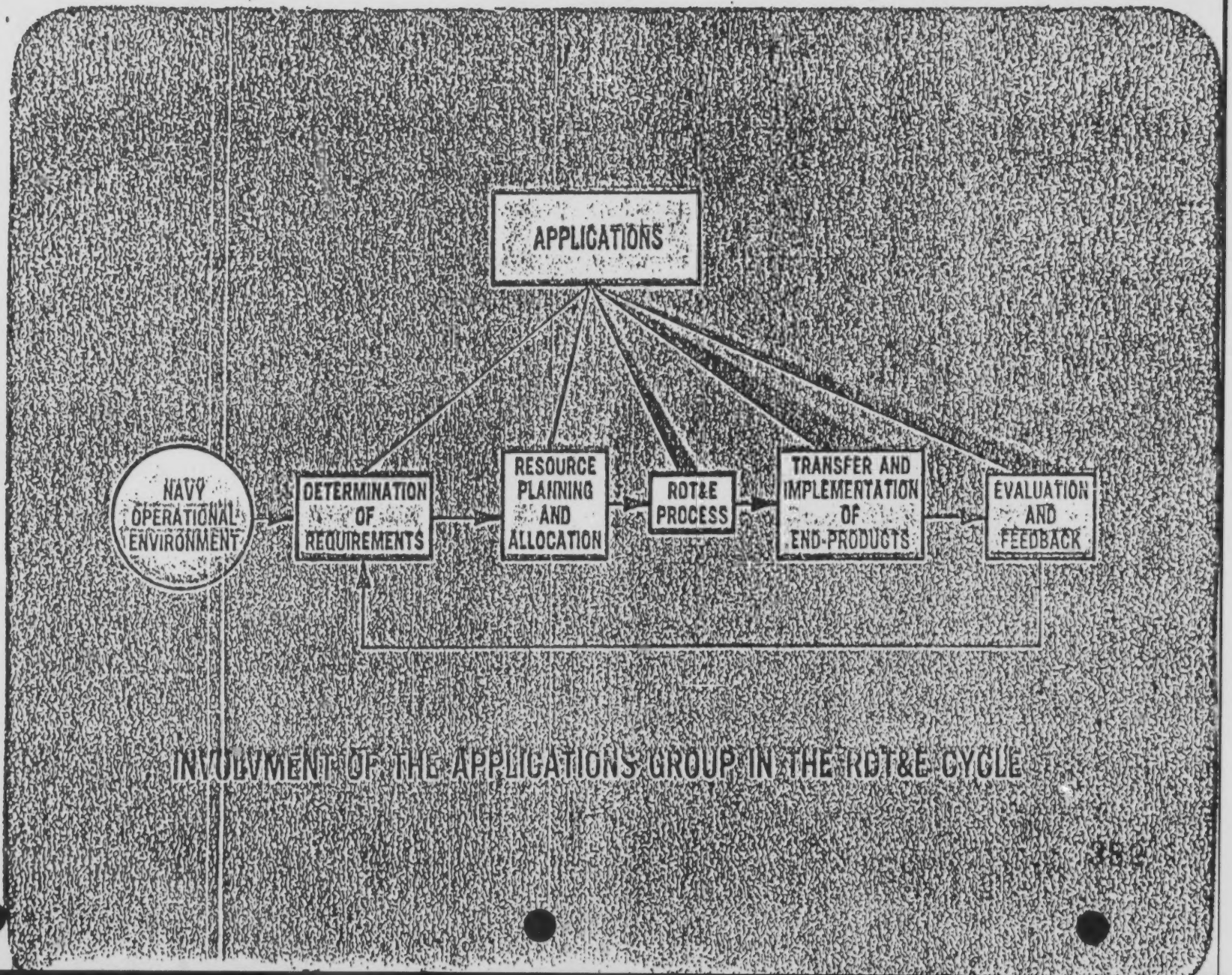
- INADEQUATE COMMUNICATION OF IMPORTANT PROGRESS
- TOTAL DOLLAR INVESTMENT IS TOO SMALL
- GENERAL RESEARCH APPLICATION PROBLEMS
- INADEQUATE IN-HOUSE RDT&E FOLLOW-ON
- RESEARCH JUDGED USEFUL, BUT NOT USED

APPLICATION PROGRAM OBJECTIVES

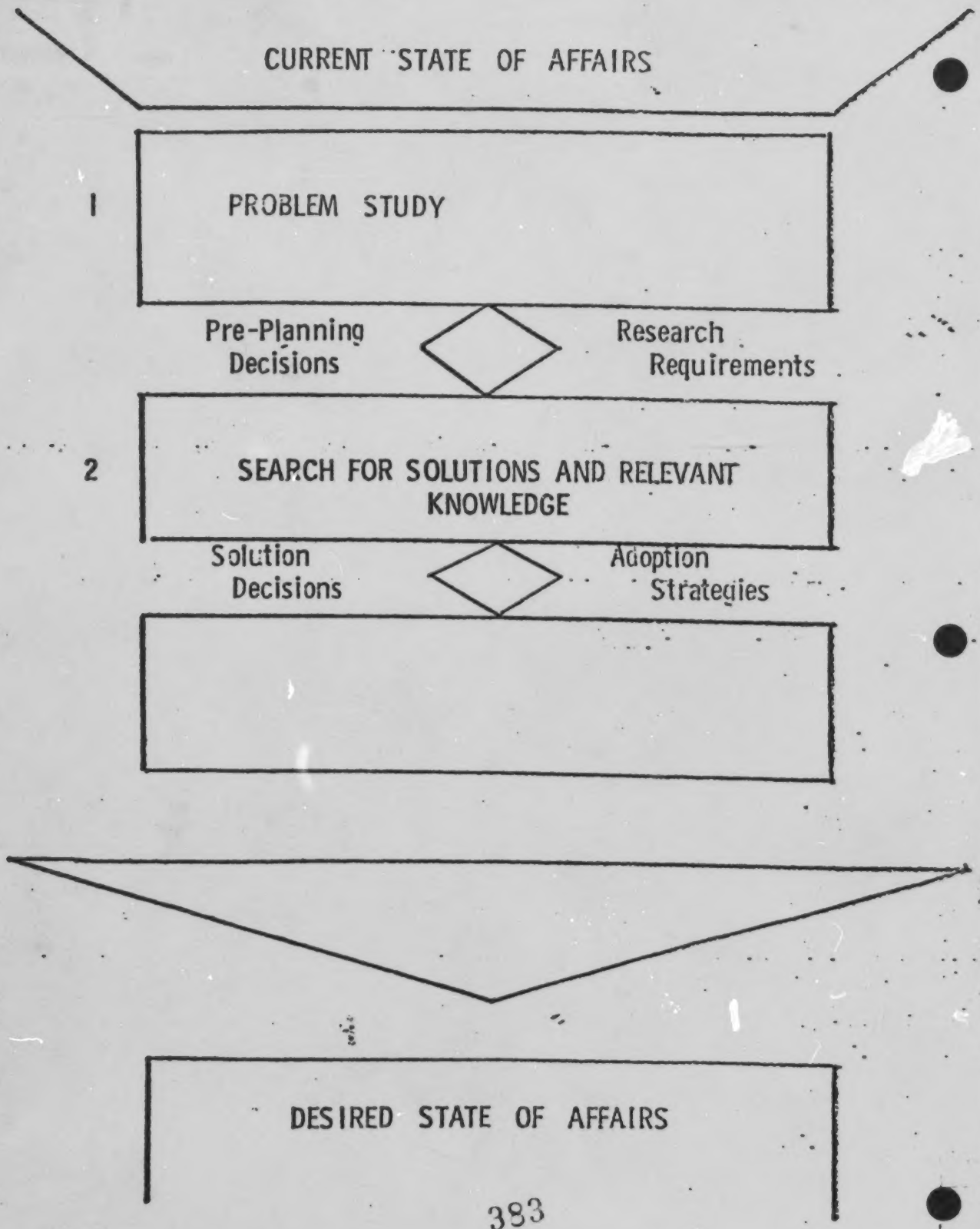
- DEVELOP MORE EFFECTIVE PROCEDURES AND TECHNIQUES TO FACILITATE THE INTRODUCTION AND USE OF NAVPERSRANDCEN RDT&E END PRODUCTS BY OPERATING FORCES OF THE NAVY.

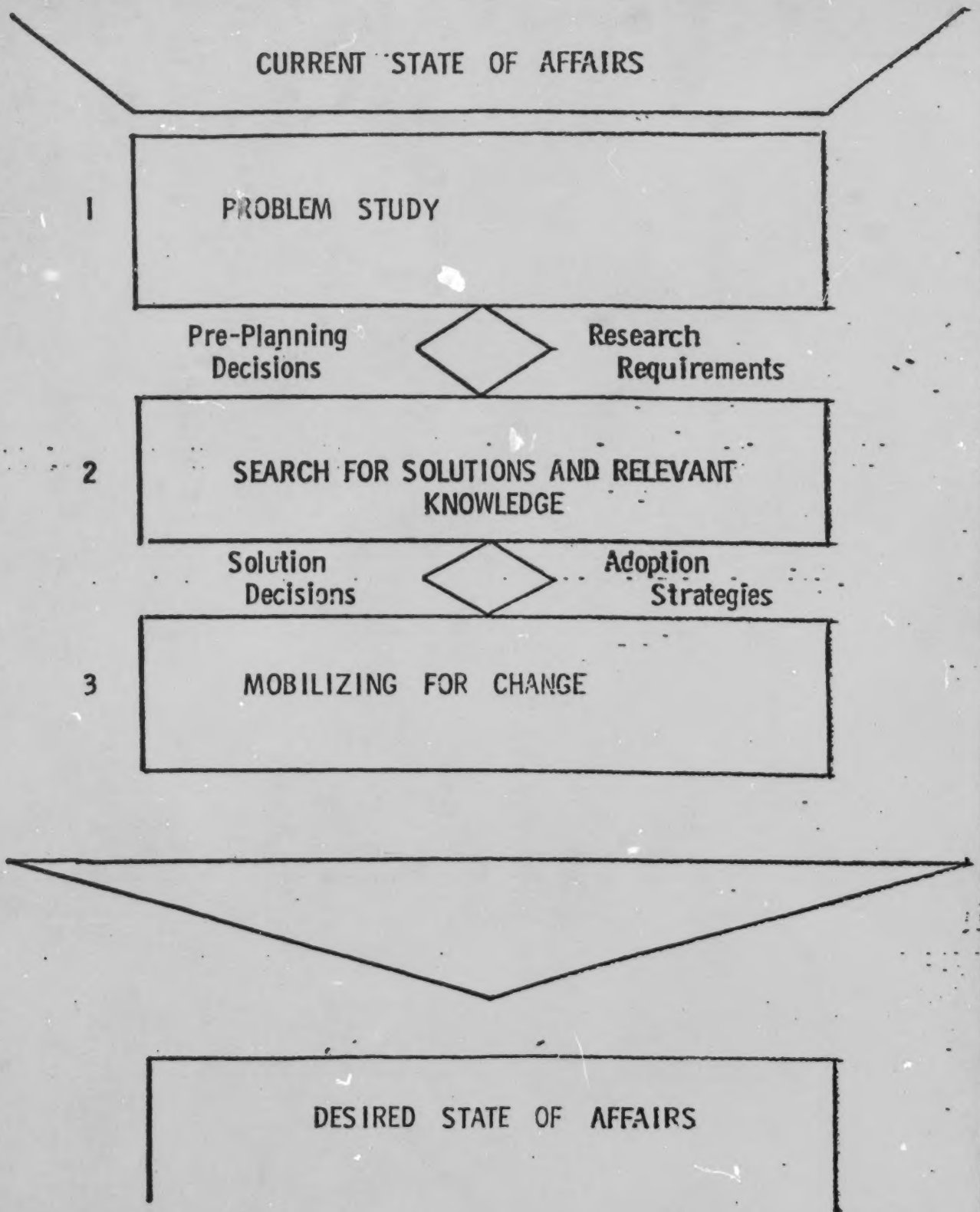
- DEVELOP AND IMPLEMENT AN APPLICATION MODEL CONSISTING OF COMMUNICATION AND COORDINATION NETWORKS CAPABLE OF DETERMINING AND ASSESSING RDT&E END PRODUCT UTILIZATION.

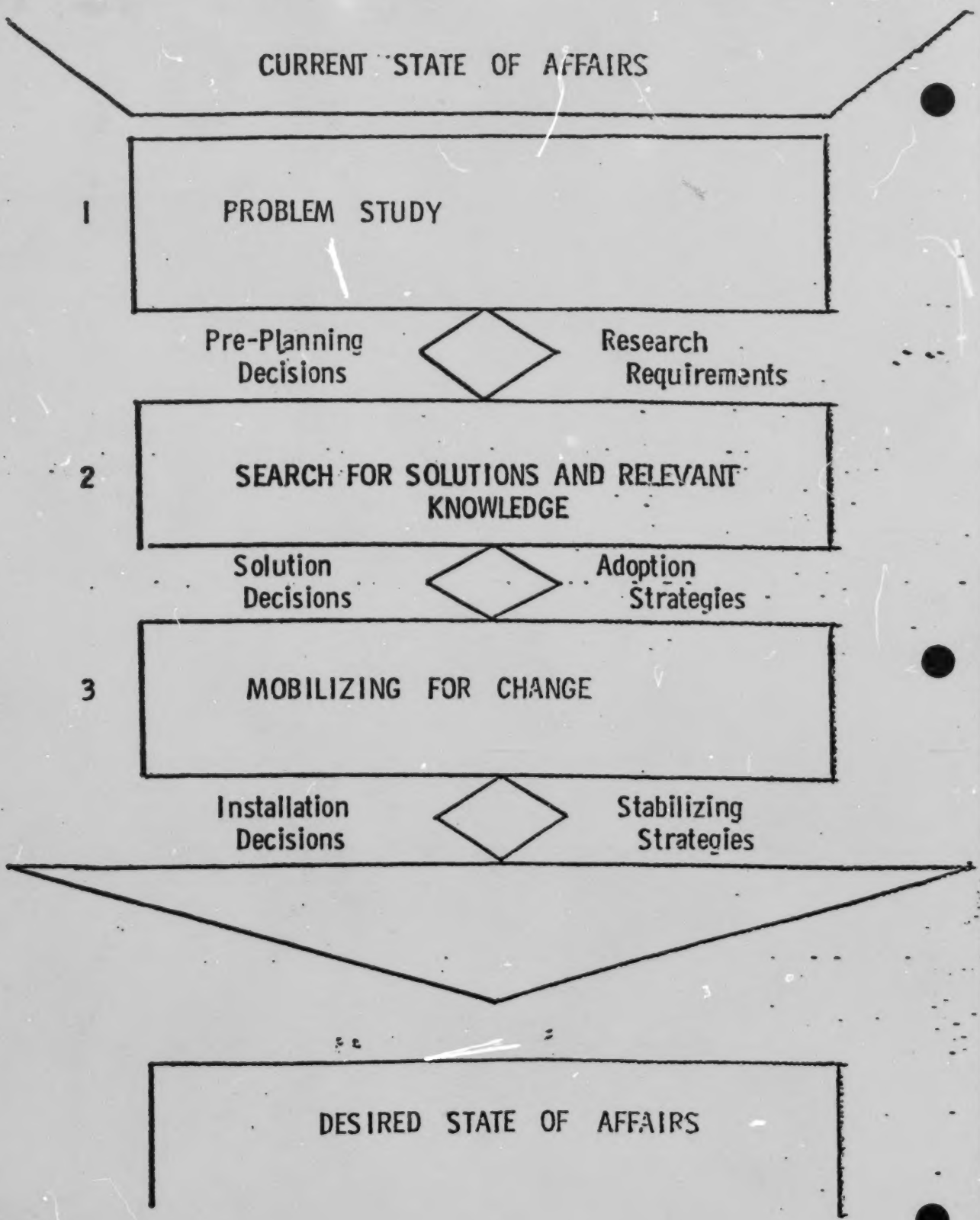
- CONDUCT STUDIES TO DEVELOP NEW METHODS AND TECHNIQUES FOR RDT&E DISSEMINATION AND UTILIZATION.

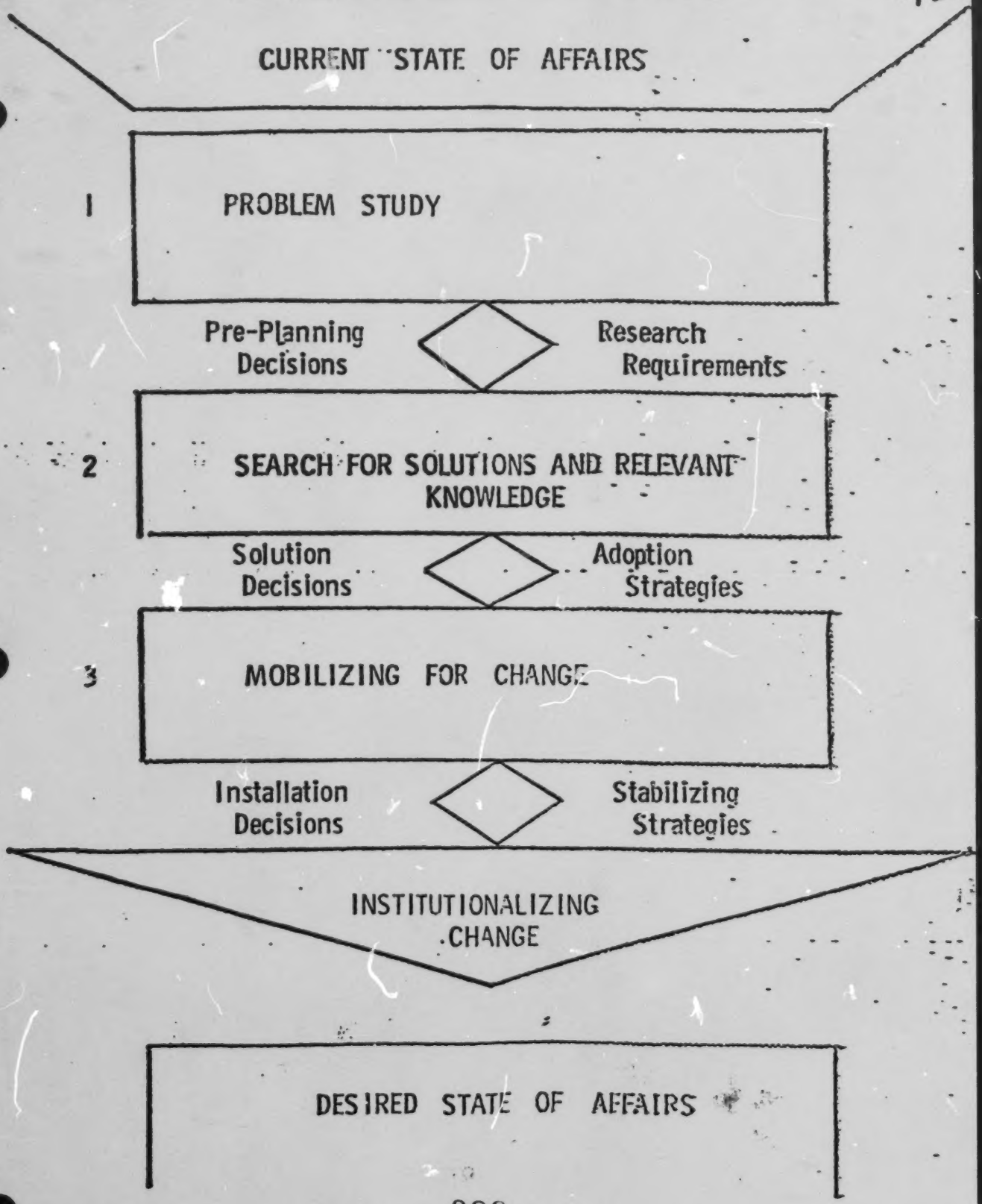


INVOLVMENT OF THE APPLICATIONS GROUP IN THE RDT&E CYCLE


CURRENT STATE OF AFFAIRS
1
PROBLEM STUDY
Pre-Planning Decisions
Research Requirements
2
SEARCH FOR SOLUTIONS AND RELEVANT KNOWLEDGE
Solution Decisions
Adoption Strategies
DESIRED STATE OF AFFAIRS


CURRENT STATE OF AFFAIRS
1
PROBLEM STUDY
Pre-Planning Decisions
Research Requirements
2
SEARCH FOR SOLUTIONS AND RELEVANT KNOWLEDGE
Solution Decisions
Adoption Strategies
3
MOBILIZING FOR CHANGE
DESIRED STATE OF AFFAIRS


CURRENT STATE OF AFFAIRS
1
PROBLEM STUDY
Pre-Planning Decisions
Research Requirements
2
SEARCH FOR SOLUTIONS AND RELEVANT KNOWLEDGE
Solution Decisions
Adoption Strategies
3
MOBILIZING FOR CHANGE
Installation Decisions
Stabilizing Strategies
DESIRED STATE OF AFFAIRS


CURRENT STATE OF AFFAIRS
1
PROBLEM STUDY
Pre-Planning Decisions
Research Requirements
2
SEARCH FOR SOLUTIONS AND RELEVANT KNOWLEDGE
Solution Decisions
Adoption Strategies
3
MOBILIZING FOR CHANGE
Installation Decisions
Stabilizing Strategies
INSTITUTIONALIZING CHANGE
DESIRED STATE OF AFFAIRS

EVALUATION OF OUTCOMES OF ATTEMPTS TO INSTITUTIONALIZE CHANGE

By Joseph Ward and Gail T. Rayner

Introduction

Efforts to institutionalize change occur with some frequency as the Army attempts to maintain an effective defense posture in an era of ever increasing costs. There is a constant upgrading of weapons systems, continuous changes in personnel requirements, and frequent modifications in doctrine and regulations. These are visible weapons of our enemies. It may be that hardware is visible that change is seen to take place--whereas training doesn't show. There is also a constant search for innovations that promise to increase the quality or reduce the cost of training or reduce the time required for training. One innovation, applying a systems approach to training (SAT), has been shown to be effective after intensive field trials. And, although SAT has been proven to be a viable tool, institutionalizing it to realize its promise has become a problem. The history of these attempts tell of many failures. If the lessons learned from some of these attempts are heeded, they can provide valuable source of information on the problem of institutionalizing an innovation in an existing system. There is documentation of several attempts to institutionalize a systems approach to training in the Army.

Development of CONARC Reg 350-100-1

In 1968, USCONARC (currently TRADOC) attempted to institutionalize a model to change the way in which its instructional programs were being developed. This instructional development model was called Systems Engineering of Training (Course Design) and was issued as CONARC Regulation 350-100-1. This regulation established the procedures for developing instructional programs which would relate them more closely to entry level job performance requirements. These procedures had evolved over the years as a result of Army research and development efforts on a wide variety of programs. One of the first descriptions of the systems engineering model was by Crawford in 1962. He advocated a seven-step model of program development whose steps, although sparse, could be followed by experienced developers of instructional programs. The seven steps were: 1) analysis of the operational subsystem, 2) analysis of the job, 3) specification of knowledges and skills, 4) determination of training objectives, 5) construction of the training program, 6) development of measures of job proficiency, and 7) evaluation of the training program. Subsequent models developed by the Army include all of these steps, except the first--analysis of the system--a step which is seldom, if ever, accomplished.

There were two problems with Crawford's model. The first was that it was incomplete in terms of the procedures utilized and the second was the lack of detail necessary for the accomplishment of the tasks. Only expert and experienced developers were able to fill in the detail omitted in Crawford's article. However, most instructional program development was accomplished by school instructors whose knowledge and expertise were insufficient to use the Crawford model.

CONARC Reg 350-100-1 was developed essentially as a handbook that could be used by instructors with limited developmental experience. Its introduction required that training programs be designed in a substantially different manner than previous development efforts. It arrived with suspense dates that almost precluded human effort as a means for meeting the Reg's requirements. It also had an unwieldly control system which included sending Job Task Data Cards (JTDC) to CONARC for approval. There were problems inherent in using the JTDCs, since they were often developed from existing courses rather than any real job analyses, and once they were developed it seemed as if no other substantive changes had to be made. The most complete studies of the implementation of 350-100-1 were the attempts at Ft. Rucker.

Evaluation of Attempts to Institutionalize CONARC Reg 350-100-1 Ft. Rucker, 1968.

CONARC Reg 350-100-1 represented a substantial change from the way in which the Army had been developing instructional programs. As a result, some, but not all, difficulties were anticipated in its implementation. This effort might be considered as a classic example of failure to institutionalize change by attempting to regulate it, and of a failure to understand the implications of implementing change.

Some of the reasons for this implementation failure were:

1. The model assumed a high level of expertise in Systems Engineering Technology that was not available in the TRADOC CONARC system at that time. The regulation was too difficult for personnel with limited skills and knowledge and proper guidance for this level personnel was not provided. In addition, the description of sub-products was incomplete.

2. The regulation failed to consider the coordination necessary for development of a complete training program. Complete training involves unit and school. Funds and authority necessary for coordinating with units on establishing training requirements were not available. Training requirements were established with school personnel only.

3. Failure to realize the necessity for strong, knowledgeable, centralized management to monitor progress and obtain cooperation from conflicting elements made for many problems in implementation.

4. There was also a failure to consider the time requirements of Systems Engineering and results were expected immediately. Many schools required instructors to apply Systems Engineering while carrying out regular instructional duties.

FEEDBACK RELATIONSHIPS

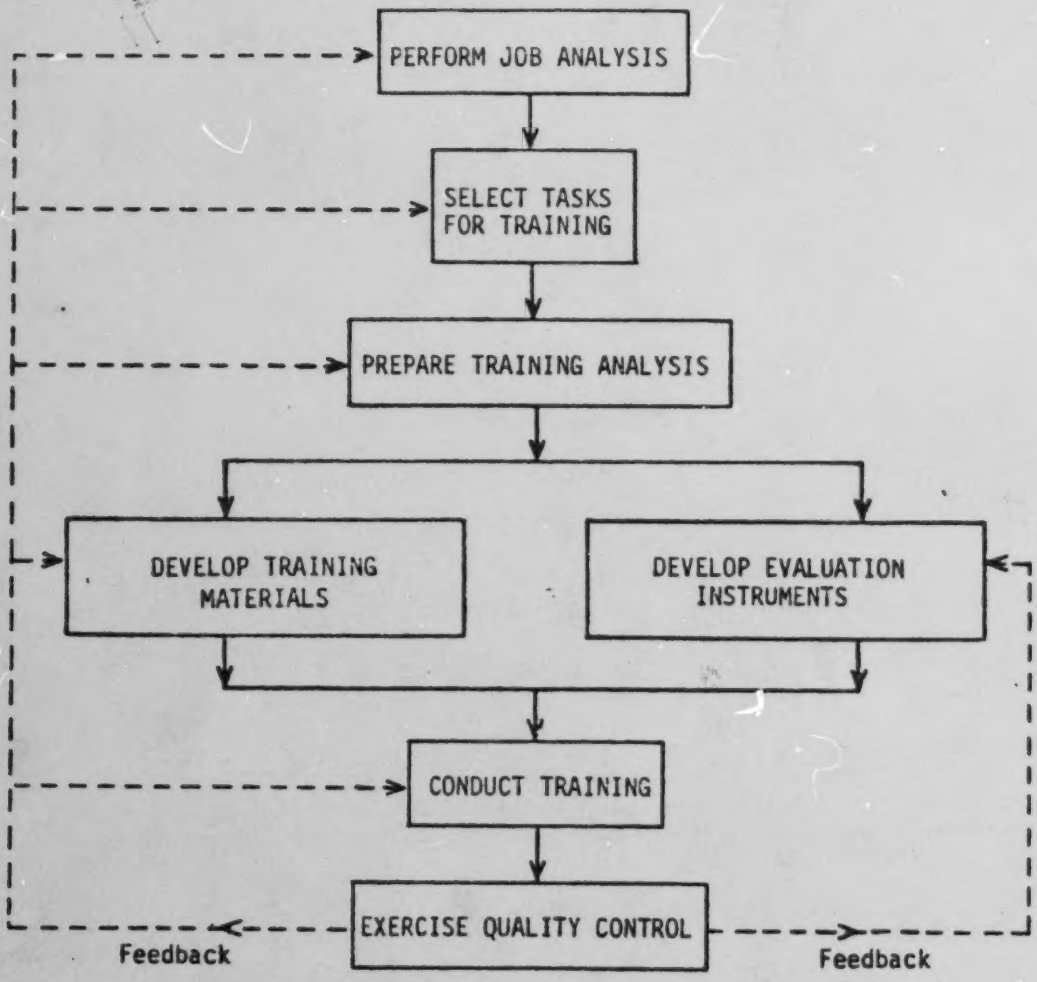


Exhibit 1: CON Reg 350-100-1

A more recent attempt to institutionalize a systems approach was the effort this year at Ft. Gordon.

Institutionalize Interservice ISD
Ft. Gordon, 1975

Even though this major effort at institutionalizing Systems Engineering had not been successful, there were schools in which personnel were utilizing Systems Engineering very effectively. The implementation problems and the isolated successes of 350-100-1 were studied prior to efforts to institutionalize the Interservice Instructional Systems Development Model at the U.S. Army Signal School at Ft. Gordon, Georgia. The ISD procedures had already been tested on Army and Navy military and civilian personnel, and revisions had been made based on these trials. The Ft. Gordon effort was purely a trial at implementing the procedures.

Some of the steps taken in the trail implementation were:

1. The selection of a school based on projected ability to perform and accept change. This was neither the largest TRADOC school nor the smallest.

2. People to manage the effort were identified at all levels.

3. Command briefings were held to get support and permission.

4. The project was approved at senior 0-6 level.

5. Approval and support at TRADOC was secured.

6. The managers and technical level people were trained in ISD and also trained to train all personnel in the project.

However, even good plans cannot guarantee success and this implementation attempt was also unsuccessful. Although there was support and expertise at the Signal School, there was a strong will at some critical levels to contain ISD and prevent it from "tainting" the school. This negative attitude toward change greatly hindered the implementation process and eventually stopped it. However, there is still a core of personnel who can Systems engineer introduction at the school.

Ongoing Effort to Institutionalize Interservice ISD

The Navy has used the most systematic approach in its ISD implementation. The control of the implementation effort is at a very high level and the training of the technical level personnel is centrally quality controlled and centrally administered. A team approach is used, with specialists performing specific tasks. The training procedures have been modified as needed to meet the specific needs of the service providing increased relevance for implementation and its success.

The Navy attempt is currently on-going and the probability of success seems high.

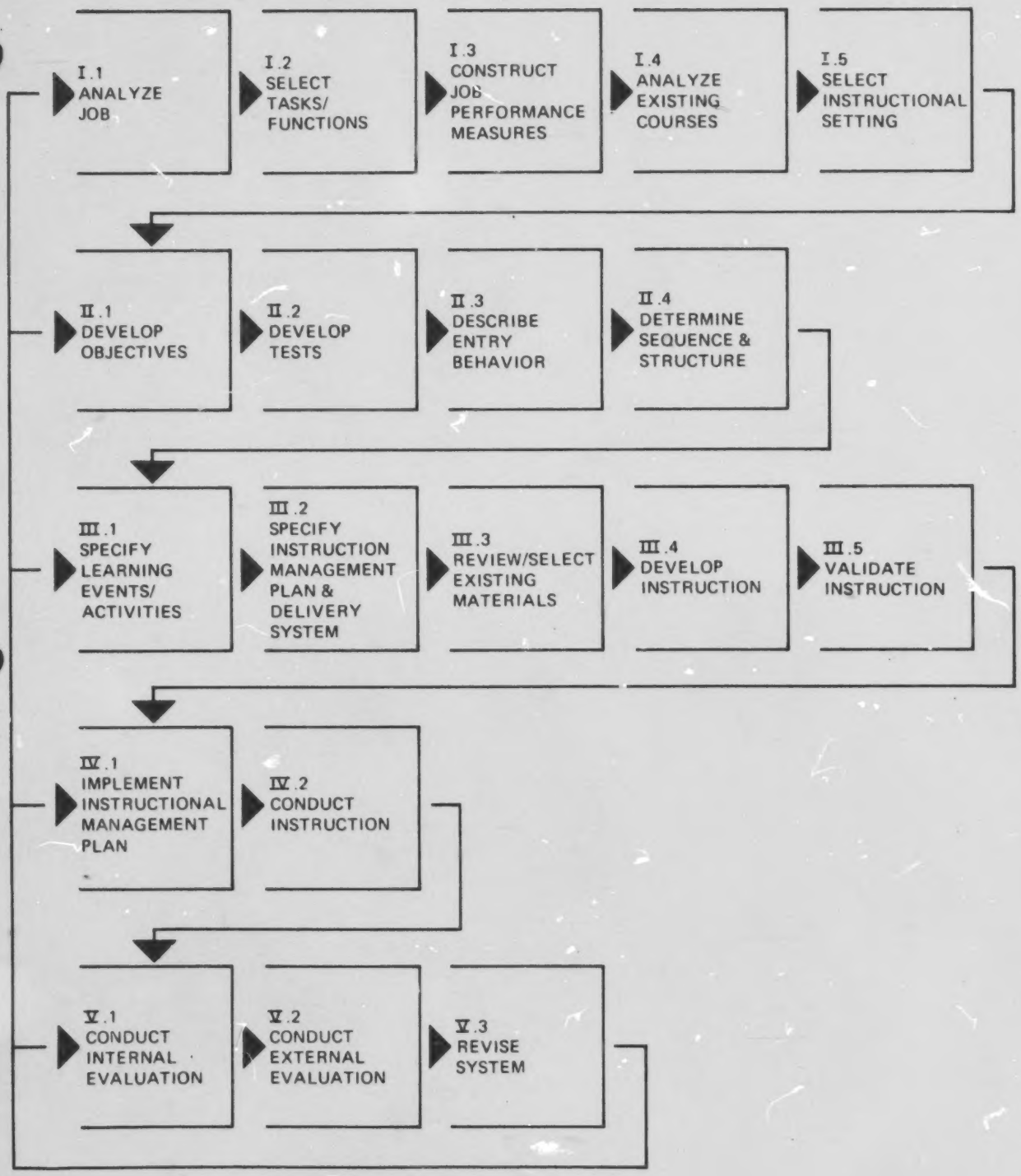

I.1 ANALYZE JOB
I.2 SELECT TASKS/ FUNCTIONS
I.3 CONSTRUCT JOB PERFORMANCE MEASURES
I.4 ANALYZE EXISTING COURSES
I.5 SELECT INSTRUCTIONAL SETTING
II.1 DEVELOP OBJECTIVES
II.2 DEVELOP TESTS
II.3 DESCRIBE ENTRY BEHAVIOR
II.4 DETERMINE SEQUENCE & STRUCTURE
III.1 SPECIFY LEARNING EVENTS/ ACTIVITIES
III.2 SPECIFY INSTRUCTION MANAGEMENT PLAN & DELIVERY SYSTEM
III.3 REVIEW/SELECT EXISTING MATERIALS
III.4 DEVELOP INSTRUCTION
III.5 VALIDATE INSTRUCTION
IV.1 IMPLEMENT INSTRUCTIONAL MANAGEMENT PLAN
IV.2 CONDUCT INSTRUCTION
V.1 CONDUCT INTERNAL EVALUATION
V.2 CONDUCT EXTERNAL EVALUATION
V.3 REVISE SYSTEM

Exhibit 2: Interservice ISD Model

CONARC Reg 350-100-1 was developed essentially as a handbook that could be used by instructors with limited developmental experience. Its introduction required that training programs be designed in a substantially different manner than previous development efforts. It arrived with suspense dates that almost precluded human effort as a means for meeting the Reg's requirements. It also had an unwieldly control system which included sending Job Task Data Cards (JTDC) to CONARC for approval. There were problems inherent in using the JTDCs, since they were often developed from existing courses rather than any real job analyses, and once they were developed it seemed as if no other substantive changes had to be made. The most complete studies of the implementation of 350-100-1 were the attempts at Ft. Rucker.

Evaluation of Attempts to Institutionalize CONARC Reg 350-100-1 Ft. Rucker, 1968.

CONARC Reg 350-100-1 represented a substantial change from the way in which the Army had been developing instructional programs. As a result, some, but not all, difficulties were anticipated in its implementation. This effort might be considered as a classic example of failure to institutionalize change by attempting to regulate it, and of a failure to understand the implications of implementing change.

Some of the reasons for this implementation failure were:

1. The model assumed a high level of expertise in Systems Engineering Technology that was not available in the TRADOC CONARC system at that time. The regulation was too difficult for personnel with limited skills and knowledge and proper guidance for this level personnel was not provided. In addition, the description of sub-products was incomplete.

2. The regulation failed to consider the coordination necessary for development of a complete training program. Complete training involves unit and school. Funds and authority necessary for coordinating with units on establishing training requirements were not available. Training requirements were established with school personnel only.

3. Failure to realize the necessity for strong, knowledgeable, centralized management to monitor progress and obtain cooperation from conflicting elements made for many problems in implementation.

4. There was also a failure to consider the time requirements of Systems Engineering and results were expected immediately. Many schools required instructors to apply Systems Engineering while carrying out regular instructional duties.

Speculation about the Future of a Systems Approach in the Army

Mistakes were made in implementation attempts of Systems Engineering and ISD in the Army. There are still high level influential people who do not want a systems approach--but whatever you call it or however you arrange the boxes or draw the arrows, it's going to happen. There is a critical mass of support plus the means to insure technical expertise. Whether or not the Systems Engineering or ISD approach is used to develop the instruction, the product of the Army Schools is supposed to be a trained men. With job analysis data provided by MORD and with SQTs used to evaluate the personnel in the field, good quality control of training should be possible. It will be up to the schools to successfully train the individual no matter what method they use to develop their training. But only a systems approach will give them that trained man efficiently.

REFERENCES

Crawford, M.P. Concepts of training. In R. M. Gagne (Ed.), Psychological principals in systems development (2nd ed.). New York: Holt, Rinehart and Winston, 1966.

Ricketson, D. S., Schulz, R. E., and Wright, R. H. Review of the CONARC systems engineering of training program and its implementation at the United States Army Aviation School. Ft. Rucker, AL. Human Resources Research Organization, April 1970.

STRATEGIES FOR INSTITUTIONALIZING CHANGE

R. M. Morgan and Unna Huh Jeon[1]

Introduction

Institutions without the means for change and renewal cannot improve. In order to survive in this increasingly change-oriented society, any organization must be keenly aware of the existence of new knowledge and be able to use this new knowledge. This awareness leads to discussion then to appraisal and finally it may lead members of the organization to feel a need for change within the organization itself, and, thus, begins the process of innovation.

Yet, despite the fact that a need for change and innovation may be felt by certain members of the organization, it often happens that innovation, if it occurs at all is superficial. For effective innovation to occur, the organization must have been ready for it, both psychologically and structurally. There are, however, many variables which may inhibit or prevent adoption of change within an organization.

Inhibitors

1. Psychological Inhibitors

Certain psychological conditions within an organization will often inhibit the introduction of innovations. A desire to maintain organizational stability, fear of new knowledge and simple complacency are such conditions.

Stability

A major barrier to the acceptance of new knowledge is the desire on the part of organization members to maintain the stability of relationships

---

[1] Morgan is Director of the Instructional Systems Development Center, Florida State University. Dr. Jeon is a Research Associate in ISDC.

within the organization and with other groups. Entry of an innovation is blocked when members of the organization perceive that the advantage of accepting new knowledge is outweighed by the possible disadvantages resulting from disturbing this stability. The desire for stability is an overriding concern of some members of all organizations and results in resistance to change.

Fear of New Knowledge

New knowledge can also be viewed as a threat to individual members of the organization. Because of the possible implication of individual failure or incompetency that new ideas from the outside might bring with them, the defenses of these individuals rise like a Chinese Wall to bar the entry of new ideas. This may be the source of the NIH (Not Invented Here) syndrome--if the idea or process was not internally generated it couldn't possibly be appropriate for the local organization.

Complacency

Often when an organization has achieved some measure of initial success, complacency follows. Once complacency develops within the organization its members frown through jaundiced eyes at the suggestion of change, having already perceived that change is unnecessary.

2. Structural Inhibitors

The structure of the system also exerts a strong inhibiting influence upon the adoption of new ideas. Some major inhibiting properties associated with organizational structure may be summarized as hierarchy, lack of long-range planning, unique systems norms and pattern of reward.

Hierarchy

Military systems are characterized by a relatively rigid organizational hierarchy. Many observers believe that in the military, since change may be ordered from above, the adoption of new ideas in the military system must necessarily be prompt and efficient. However, several students of organizational behavior have suggested that the more hierarchical the structure of

the organization, the less the possibilities of change (Havelock, 1976; Rogers, 1971).

Hierarchical structure has two shortcomings in the adoption of change: (1) message distortion and (2) lack of feedback. First, as the information flows down the hierarchy, "progressive segregation" occurs (Griffith, 1964). Messages become selectively diluted. Consequently, the message received at the bottom of the hierarchy is poor and incomplete, leading to poor and incomplete behavioral change. Second, because almost all the information flow is downward within the hierarchy, organization members are reluctant to communicate upward unless the message reflects posivie evaluation of themselves. They may also be ineligible to communicate upward. Therefore, behavioral changes of organization members in the hierarchy often reflect a process of "compliance" in contrast to a process of "introduction" and "internationalization" (Havelock, 1976). The adoption of change directed by others may turn out to be both superficial and unstable and hence not likely to be permanent.

ck of ng-range anning

For successful adoption of change, long-range planning is prerequisite. The time dimension involved in the innovation-decision process, the innovativeness of the effected individuals and the rate of adoption of an innovation in a system all have to be carefully considered. Change efforts without long-term planning are likely to be ineffective because members of the organization are pushed to adopt change prematurely. The high mobility of personnel in the military makes the implementation of long-range planning more difficult.

stem Norm

Systems norms, the established behavioral pattern for the members of a given system (Rogers, 1971, p. 31) also can be a barrier to change. Systems norms define a range of acceptable behavior and serve as a guide

for the members of the system. They act either as incentives or restraints on individual decision-making either to adopt or reject change. When system norms are not favorable for change, a decision to adopt change is difficult for organization members to make.

Military systems tend to develop unique system norms of their own. A common coding system or an esoteric language only relevant within the military are examples of a military system norm. These unique system norms make it difficult for the system to communicate with outsiders. What Deutsch (1963) labeled "communication differential" occurs. This fosters a lack of understanidng of new information from outside the system and accordingly results in small motive for chaning the traditional ways of doing things.

Reward Pattern

The pattern of reward is a key factor in motivating organization members to adopt change. When a system rewards its members only for their stable and habitual behavior--and this is typical in most organizations, including military--innovative activities are likely to be discouraged.

Possible Solutions

Now we shall discuss some basic strategies that may be used to help overcome these psychological and structural inhibitors of change.

Linkage

One strategy is to overcome the lack of communication with external information resources and ignorance of existing outside knowledge. This lack of information with its consequent ignorance results in a basic psychological unreadiness for the introduction of change. Awareness of outside knowledge can be gained through the establishment of strong linkages with outside resource systems (Havelock, 1976). One useful way of establishing these linkages is the use of a change agent from outside the organization. An outside change agent who is familiar with both outside resources and

a. Change Agent

internal problems of the organization can form a crucial link between the organization and outside information resources. Initially the change agent should help make members of the organization aware of the need for change.

The change agent suggests new alternatives, dramatizes existing problems and convinces people that they are capable of confronting these problems. Establishing rapport with an outside change agent may reduce the resistance to change that is based on unawareness of alternatives, allowing change to proceed.

. Internal Subunit

Still another way of linking an organization with external knowledge sources is the formation within the organization of a subunit designed to retrieve, sift and adopt ideas from the outside. The ideal functions of this subunit would be to: (1) sense the change in the environment, (2) determine the need for change in the organization, (3) seek outside and/or internal information corresponding to needs, (4) identify suitable alternative approaches, (5) assist in the implementation of the selected alternatives, and finally (6) evaluate the innovation.

ıange riented ɔperative ɔirit

Resistance to change based on fear of new knowledge and deriving from system norms unfavorable to change can be diminished by encouraging a change oriented, cooperative spirit among the organization members. Active participation in the process of change of those members who are responsible for the implementation of change and who make the decision whether to adopt or reject innovation is crucial. This participation leads them to the identification of needs and to the investigation of different approaches that may be responsive to these needs. A sense of shared participation in the definition and solution has been shown to significantly effect people's motivation to change.

. Partici-pation

. Training

Training of top and middle managers to value and reward innovative

activities and the exchange of knowledge among their personnel also will create an orientation toward change. Personnel training and development programs which concentrate on the importance of being aware of new knowledge, the basic problems of knowledge flow, ways to improve knowledge flow through interpersonal linkages, establishment of new goals for the organization and implementation of planned change toward new goals, can be very useful.

c. Change of Leadership

The recruitment or employment of staff with a potential for innovative actions, or a change of leadership in the organization can sometimes bring about an overall change-oriented climate within the organization. This climate can encourage and facilitate innovative activities.

Experimentation

Change is also facilitated by experimentation. Experimentation allows the members of an organization to gradually familiarize themselves with an innovation and diminish their fear or doubt about the innovation prior to a continuing commitment to it. It also provides them with an opportunity to evaluate for themselves the effects of a proposed change. Dissemination of the results of experimentation with a new idea serves as a basis for adopting it. Successful trials of an innovation can lead to its being adopted permanently. Top organizational administrators should foster a scientific attitude toward experimentation and seek ways to support the implementation staff in experimentation with new ideas.

Communication Flow

As has been suggested previously the military organization is particularly susceptible to message distortion and lack of feedback in the transmission of new ideas because of the prevailing downward flow of communication. To overcome these pitfalls within the existing hierarchical structure, a free flow of communication should be facilitates through the peer-recognized opinion leaders (Havelock, 1976; Rogers, 1971). Opinion leaders are individuals characterized by being somehwat more cosmopolitan

and flexible, therefore having greater access to outside information resources or change agents. They have to be accessible to their fellow members through social or informal face-to-face discussions. Opinion leaders can serve as an effective communication link within the hierarchy: delivering a message of change from above to all members, and communicating upward feedback from the operation level. Arndt indicated the influential power of opinion leaders by reporting: "Those receiving favorable word-of-mouth communication (from opinion leaders) were three times as likely to buy the new product as were those receiving unfavorable word-of-mouth" (Arndt, 1968). In a study of educational change in ten Latin American countries, we are just concluding at Florida State University, we have found empirical evidence that indicates, at least in developing countries, "word-of-mouth" communication is by far the most potent--if not the only--impetus to educational change.

By enlisting the cooperation of organizational opinion leaders, the efficiency of change efforts can be increased, and the development of possible antagonism toward innovation diminished. To secure the support of influential opion leaders, top administrators must accurately identify them, and insure their participation in the decision-making and implementation process.

Mass Communication

Another method of effective information transfer within the hierarchy is the use of mass communication channels. The interpersonal communication method through opinion leaders that has been mentioned is the most persuasive in getting individuals to adopt new ideas. However, in actually conveying new knowledge, the mass media are more effective in increasing awareness of and the accurate transfer of the original message. Therefore, effective selection and utilization of mass communication channels should be encouraged.

Long-range Planning

a. Change Agent

Problems in the implementation of long-range change planning caused by high mobility of personnel within the military system can be partially alleviated by the use of a change agent from outside the system. A successful change agent should be able to diagnose a client's need from the view point of the client and creatively search among alternatives for the best solution compatible with the client's needs. The change agent makes a long-range plan for adoption of the solution considering the receptivity to innovativeness of the target population and the present condition or readiness of the client system. Once a long-range plan has been made and accepted by the client system, the change agent may need to continue to devote time to the implementation and renewal process of planned change in order to help institutionalize change. To maintain an effective relationship with the change agent and in turn bring about the successful institutionalization of change, top administrators should clearly define the role of the change agent and provide active support while his staff members should closely work with the change agent.

Reward

Finally, the system has to build an internal strategy to provide consistant rewards for innovative activities. Perceived relative advantage or a past experience of reward for utilization efforts are just as important as actual present rewards. Whatever reward is most meaningful to the members of an organization should be used as a reinforcing mechanism if the meximum institutionalization of change is to be achieved.

BIBLIOGRAPHY

Amatea, E.S. Initiating the diffusion of the elementary career development guide, the yellow brick road. Final Report, Project No. VTAD-D5-005, University of Florida, Gainesville, Florida, 1975.

Arndt, J. A test of the two step flow in diffusion of a new product. Journalism Quarterly, 1968, 45, 457-465.

Deutsch, K.W. The nerves of government: Models of political communication and control. New York: Free Press of Glencoe, 1963.

Dodl, N.R., Kibler, R.J., & Dick, W., et. al. The Florida assessment and diffusion system model. Florida State University, December, 1974.

Griffiths, D.E. Administrative theory and change in organizations. In M.B. Miles (Ed.), Innovation in education. New York: Bureau of Publications, Teachers College, Columbia University, 1964.

Guba, E.G. Diffusion of innovations. Educational Leadership, 1968, 25, 292-295.

Havelock, R.G. The change agent's guide to innovation in education. Englewood Cliffs, N.J.: Educational Technology Publications, 1973.

Havelock, R.G. in collaboration with Guskin, A. Planning for innovation through the dissemination and utilization of knowledge. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan, 1976.

Rogers, E.M., & Shoemaker, F. Communication of innovations: Across-Cultural Approach. New York: Free Press of Glencoe, 1971.

Rogers, E.M., & Svenning, L. Managing change. Report to OPERATION PEP: A Statewide project to prepare planners for California, USOE/DHEW Title III grant, San Mateo County Board of Education, Burlingame, California, September, 1969.

"THE OTHER TEN PERCENT" -- REVISITED

J. E. Gerber, Jr., Ph. D.
United States Army Forces Command
Fort McPherson, Georgia 30330

Presented to the 18th Annual Conference of the
Military Testing Association,
Gulf Shores, Alabama,
18-22 October 1976

The opinions and conclusions expressed in this paper are the author's and are not necessarily those of the United States Army or of Forces Command.

ABSTRACT

This paper gives an eleven year perspective to the development of an emerging theory of individual learning and criterion testing. It reviews the author's paper titled "The Other Ten Percent" delivered to the third annual convention of the National Society for Programmed Instruction, Philadelphia, May 1965. That paper focused on those learners who were left out of the then popular criterion for program effectiveness -- success on 90% of the criterion test items or frames by 90% of the learners -- the so-called 90 by 90 standard. The paper advanced some basic thoughts for construction of a model of individual learning styles and advocated use of a variety of instructional styles to develop alternate forms of programs which might be more compatible with and less biased against individual learning styles. The present paper examines some writings of the period 1965-1976 on theoretical and practical considerations in the development of self study and individual learning and testing materials, and offers some suggestions and conclusions.

In the decade of the 1960's, the Socratic principle of tutorial teaching was rediscovered with intense fervor and applied in a self study medium called programmed instruction. At that time, programmed instruction was a revolutionary concept because it broke with the traditional practice of basing the criterion for successful teaching and learning on the bell-shaped curve and substituted a much higher criterion. The standard which became popular was the so-called 90 by 90 criterion. That is, if 90 percent of the target population studying the program achieved success on 90 percent of the criterion frames or test items, the self instruction program was considered a successful teaching instrument.

At the May, 1965, convention of the National Society for Programmed Instruction held in Philadelphia, I presented a paper titled "The Other Ten Percent" which took cognizance of and applauded the 90 by 90 criterion, then asked, "What about the other ten percent?" What about the students who fail to achieve success on 90 percent of the items? The paper advanced the thought that at least part of the reason people fail to demonstrate criterion performance on a self teaching instrument is that the instrument or program is cast in a

teaching mode or style that is incompatible with or biased against the individual's learning style. To make the concept more concrete, I postulated an arbitrary set of six pure form learning styles, identified by letters A through F. I chose the number six because permutations of it are easily plotted on a bell curve. In actuality, perhaps the number was not entirely arbitrary, as I was influenced by the work of some of the writers of the time of whom it may be said were pioneers in an intellectual ferment preliminary to many of the innovations which are now beginning to come into being. Among these were the factor analytic works of Cattell (1948, 1949), the Guilfords (1938), French (1951, 1965), and Fruchter (1954), and the work in human learning by Tolman (1949) and Westcott (1962) among others.

Next, I assumed a teaching program or self study instrument that highly favored the student possessed of learning style A and discriminated severely against the person whose learning style was pure form style F. Expected performance is shown on this

SLIDE

taken from the 1965 paper which is attached to this one as a supplement. Here we see that the person with learning style A working on this program would achieve maximum learning in minimal time and maximum retention of content for maximum time. The student with style F attempting to learn from this same program would be expected to achieve minimally in maximum study time, and promptly lose what little he gained. The others, those whose learning styles are other than A or F, or are mixtures of styles A through F, would achieve and retain competency variously according to the compatibility or facility of their learning styles versus the bias of the program toward learning style A.

SLIDE OFF

The message of the 1965 paper to programmers was to capitalize on their break with the traditional bell curve, study individual learners as individuals, abandon the practice of writing programs for heterogeneous groups of people which only looked homogeneous, and begin writing individual teaching instruments to fit individuals.

The purpose of the present paper is to examine some of the literature of the intervening period to see how far we have moved in this direction since 1965.

It becomes immediately noticeable that even though the proceedings of the 1965 NSPI convention were not published, some of the germinative thinking expressed at that time has found its way into present day writing. Dunn and Dunn (1974), in a paper remarkably similar in concept to the 1965 paper, "The Other Ten Percent," used the basic premise that instructional programs should be matched to the particular learning style of the individual student to develop charts and suggestions for matching the two under varying instructional settings. Cawley, Miller, and Milligan (1975) found support for an hypothesis that

> In any selected population of adults, both polar analytic and polar relational, as well as mixed and conflicted, cognitive styles will be found.

Both their hypothesis and results are likewise compatible with the concept of the 1965 paper titled "The Other Ten Percent." By moving down the scale of abstraction, it may easily be seen that their "analytic" style could be a concrete application of the abstract extreme called "style A"; their "relational" style, being at the opposite end of the continuum, could be represented by "style F"; and the "mixed and conflicted styles" could fit very well on the permutations of A through F displayed on the slide.

From a slightly different approach, Johnson (1976), on the subject of evaluation of teacher competency, wrote:

> Good teaching can . . be defined in terms of a logic of potential student development, rather than in terms of any particular methods or styles which a teacher may happen to use (a matter appropriately judged only in concrete cases).

Restated, the maximally effective teaching and learning phenomenon is not a case of the teacher or programmer directing a particular set of procedures toward a group of students in the near or remote vicinity. Rather, it is a relationship between a teacher or programmer and a particular student. Friedman (1976) carried the interaction theme a step farther when he wrote:

> The questioner is just as important as the one who responds. A wise man is not a fount of knowledge. On the contrary, he is helpless until someone asks him a question great enough to evoke a profound response. He does not have the wisdom. It literally happens, comes to be, through interaction.

Under Johnson's Teacher Assessment Program (which could apply also to assessment of self study programs designed for study by individuals with individual learning styles),

> Students . . . report what they perceive the teacher to be intending, calling for, and demonstrating, and on how /the teacher/ exercises his role as an evaluator. The degree to which their perceptions approximate the self-reported goals and activities of the instructor is perceived to be one significant index of the teacher's competence.

The crux, however, is in learning more about the individual for predictive success rather than being content with after-the-fact assessment. A more specific identification of factors in individual learning styles may account, for example, for the generalized ability to learn which Madaus, Woods, and Nuttall (1974) described in their critique of Bloom's taxonomy of educational objectives. The data would be particularly revealing if it could quantify learning on a program biased against a particular individual's learning style.

Retreating from lockstep programming and advancing to administration of a series of programs of varied instructional styles on a trial-and-error basis is no solution. Translating thought into action, the American College Testing Program (1976) reported commencement of a two-year research project investigating and using a model for measuring and using students' individual learning or cognitive styles. The study reportedly will implement and evaluate a system termed in an article by O'Sullivan (1975) as "cognitive mapping" which was developed by Hill at Oakland Community College.

These, then, are a few of the papers reporting trends toward developing individual self study instruments for individuals. I have concentrated on and selected from periodicals available to and written in the language of practitioners of education. Impetus for this selection came from Hodgkinson (1976), who wrote:

> The elitist notion of several years ago that research could be done only by highly trained scholars is no longer supportable. There is now a much stronger acceptance of the proposition that the teacher, the school administrator, and even local citizens must be involved, and in fact, are likely to have some good ideas about how research can improve education.

Narrowing the selection is not intended to deprecate the effort of pure researchers building on the work of such pioneers as I have mentioned. Rather, as I wrote in an earlier paper presented to this body (1973), it is intended to help call practitioners away from lockstep. A break from lockstep is imperative, for, as Hodgkinson wrote:

> It simply isn't possible to design a single method of instruction, or a single textbook . . . that will work for all students.

Admittedly, at the other end of the spectrum, it may be equally unfeasible to design a unique teaching and testing instrument for each individual learner. But we can move, and we should move, in that direction (Forest, 1976).

Thus, the conclusion reached in the 1965 paper appears to remain valid today:

> Until programmers take into account the unique learning style of the individual, they will not have broken the lockstep, . . . leaving unanswered the question, "What about the other 10 percent?"

References

American College Testing Program. 2-year program studies learning styles. Activity, February 1976, p. 6.

Cattell, R. B. Primary personality factors in the realm of objective tests. Journal of Personality, 1948, 16, 459-487.

Cattell, R. B., and Tiner, L. G. The varieties of structural rigidity. Journal of Personality, 1949, 17, 321-341.

Cawley, R. W. V., Miller, S. A., and Milligan, J. N. Cognitive styles and the adult learner. Adult Education, 1976, 26, 101-116.

Dunn, R., and Dunn, K. Learning style as a criterion for placement in alternative programs. Phi Delta Kappan, December 1974, pp. 275-278.

Forest, L. B. Program evaluation: for reality. Adult Education, 1976, 26, 167-177.

French, J. W. The description of aptitude and achievement tests in terms of rotated factors. Psychological Monograph, 1951, No. 5.

French, J. W. The relationship of problem-solving styles to the factor composition of tests. Educational and Psychological Measurement, 1965, 25, 9-28.

Friedman, M. Aiming at the self: the paradox of encounter and the human potential movement. Journal of Humanistic Psychology, 1976, 16(2), 5-34.

Fruchter, B. Factor analysis. Princeton: Van Nostrand, 1954.

Gerber, J. E. Jr. The other ten percent. Paper presented at the convention of the National Society for Programmed Instruction, Philadelphia, May 1965.

Gerber, J. E. Jr. The growing demand for human performance testing. In Proceedings 15th annual conference of the military testing association. Brooks Air Force Base, TX: Air Force Human Resources Laboratory, 1973.

Guilford, J. P., and Guilford, R. B. Personality factors D, R, T, and A. Journal of Abnormal and Social Psychology, 1938, 34, 21-36.

Hodgkinson, H. L. Why education R&D? American Education, April 1976, pp. 11-13.

Johnson, H. C., Jr. Court, craft, and competence: a reexamination of 'teacher evaluation' procedures. Phi Delta Kappan, May 1976, pp. 606-611.

Madaus, G. F., Woods, E. M., and Nuttall, R. L. A causal model analysis of Bloom's taxonomy. American Educational Research Journal, 1974, 10, 253-262.

NSPI Journal, 1965, 4(4), 8.

O'Sullivan, K. A trainer's alphabet revisited. Training and Development Journal, 1976, 29(5), 38-47.

Tolman, E. C. There is more than one kind of learning. Psychological Review, 1949, 56, 144-155.

Westcott, M. R. Inference, guesswork, and creativity (Cooperative Research Project No. 684). Washington: U. S. Office of Education, 1962.

APPENDIX

411

THE OTHER TEN PERCENT*

J. E. Gerber, Jr.

When 90 percent of a group of students having no knowledge of a difficult subject can, after studying a program on the subject, score 90 percent or better on an achievement test on the significant points of that subject, surely the program has achieved laudable results as an instructional tool. For instructors who long may have been accustomed to thinking of achievement in terms of the bell shaped curve even to think in terms of 90 by 90 as a minimum standard of instruction represents a tremendous development in the profession of education. But for programmers to think that in the achievement of a 90 by 90 statistic they have tapped the full potential of all the learners in the target audience is to be misled by their own success. To think that 90 by 90 programs in and of themselves break the so called "lockstep" in classroom instruction is to think falsely.

The reason for this is singular, but it is manifest in two aspects: (1) ten percent of the group did not achieve criterion, and (2) programs currently are written for classes, for groups, not for individuals. While the programmer may envision the student audience as having this or that minimum background, this or that goal, or what not, there is no denying that the program is written for wide use by countless, faceless numbers. Inasmuch as any target audience of human beings is made up of individuals, the audience is not homogeneous, nor can it be made so. Accordingly, for any program having a 90 by 90 criterion, the heterogeneous audience becomes divided into two groups: (1) a heterogeneous achiever group, and (2) a much smaller and, therefore, perhaps less heterogeneous non-achiever group -- the other 10 percent.

What can be said of this non-achieving 10 percent? Why did it not achieve the magic score? Programmers, champions all of the uniqueness of the individual, cannot in good conscience maintain that people in this group do not belong in the audience. But if they do belong, what is to be said about them?

Let us examine the concept that every individual is indeed unique and entertain the thought that his uniqueness applies to his learning style just as much as it does to his physique, personality, and other attributes. Learning style,

---

*Presented to the convention of the National Society for Programmed Instruction, Philadelphia, May 1965.

as envisioned here, is that peculiar collection of attributes, conditions, and events which, when joined in a certain sequence, occasions the acquisition and perpetuation of a bit of knowledge. Whereas a full description of the theoretical basis for a learning style concept of this sort awaits development, the purpose here is merely to present the central thought for consideration in programming strategy. Let us consider that an individual who does poorly on a program with a word text, for example, might possibly devour the same content presented pictorially or graphically, or that a program appropriate for a person with high information demand might drive away the low information demander in sheer boredom. Let us concede for the moment that there may be many of these extremes of learning styles, and that those people who are in the 90 percent or achiever group are able to overcome the bias against many of them that is necessarily built into each person. There is no denying that every program has at least one bias -- the bias of instructional style. And if the programmer, as it seems likely in view of the present state of the art, has not limited his target audience to one learning style, he necessarily commits error of the first magnitude by inflicting upon the heterogeneous audience the instructional style he chooses (usually on the basis of suitability to the subject matter or personal preference).

Conceivably, such a bias might be highly facilitative to one learning style in the audience and anathema to another, with the others coping with the bias with varying degrees of success. In short, when total performance by the target audience consisting of many variants of learning styles is plotted against the four usual criteria for assessing learning from a program: -- (1) completion time, (2) percent of content assimilated, (3) percent of content retained, (4) retention time -- the bell shaped curve, so familiar to observers of composites of other naturally occurring phenomena, appears once more. But if programmers could become more selectively aware of the varied learning styles of a target audience, if they could serve the same people and, indeed, the same number of people with programs built on instructional styles compatible with the learning styles of the individuals in the audience, the results might be just as startling as those seen by the achievement of a 90 by 90 criterion.

Each learning style, as conceived here, would undoubtedly have a multiple factorial basis but would vary from the others both in the identity and relative loading of the factors. To the extend that there is a loading of common factors, there would be the ability to cope with a program instructional

style which, while effective, would not be able to bring about the maximum performance of which the heterogeneous audience is capable.

For example, suppose there are six "pure form" learning styles labeled A, B, C, D, E, and F. Suppose all of these are equipotent and non-synergic such that each acting alone or in any combination has equal power for maximum achievement. Suppose that a learner audience selected on usual criteria (such as intelligence scale score, chronological age, grade placement, and the like) studies a program built on an instructional style most favorable to Learning Style A, most discriminatory against Style F, and variously facilitative or hindering to the other four styles. Figure 1 illustrates an expected result. In the figure, those few people who have Style A in pure form complete the program in the shortest possible time, assimilate all of the criterion points, and retain knowledge of these points for the longest possible time. On the other end of the scale, those who are pure form Style F require a maximum time to assimilate little, and they promptly lose most of that. In this scheme, those of Styles F and E are the non-achievers, "the other 10 percent." But if such learning style extremes can exist, then the plight of these styles is due to instructional style bias. If this be the case, and all other variables are held constant, the overall achievement levels of Styles A and F, for example, can be reversed by reversing the instructional style bias. Admittedly, the demonstration of such a dramatic reversal in highly intelligent learners may prove difficult because of the mental resourcefullness of these people. On the other hand, it does not seem unreasonable to submit, for example, that a certain individual who achieves remarkably well through study of programs that permit use of, say, "photographic memory" or eidetic imagery may gain essentially nothing from a program that imparts information solely by, say, audible signal.

It may be seen from the figure that Styles B, C, and D achieve with varying degrees of success as they are able to cope with the instructional bias toward Style A. The figure shows the distribution of the one-way combinations of the six styles when each pure style is weighted as a standard deviation from a mean of zero in a z-score distribution. A, B, C, D, E, and F thus have weights of 3, 2, 1, -1, -2, and -3 respectively. Blank spaces under the curve may be filled with recombinations of the one-way combinations shown and with combinations of the multiple factors subsumed under each style represented by a letter. In this scheme, a weight of three represents maximum proficiency, and minus three represents

minimal proficiency. Style B, for example, has a weight of two. A style combining equally the elements of Style A with those of Style C also has a weight of two computed as follows: $(A + C)/n = (3 + 1)/2 = 2$. So does the combination ABC, in which $3 + 2 + 1 /3 = 2$. Thus, even though these three styles are of different composition, they are plotted on the same axis because they represent the same capacity for achievement on a program having an instructional style bias for Learning Style A. In other words, the program is successful with these people in spite of itself and despite the fact that it has not occasioned in Style B students (and others who are not Style A) the maximum proficiency of which they are capable.

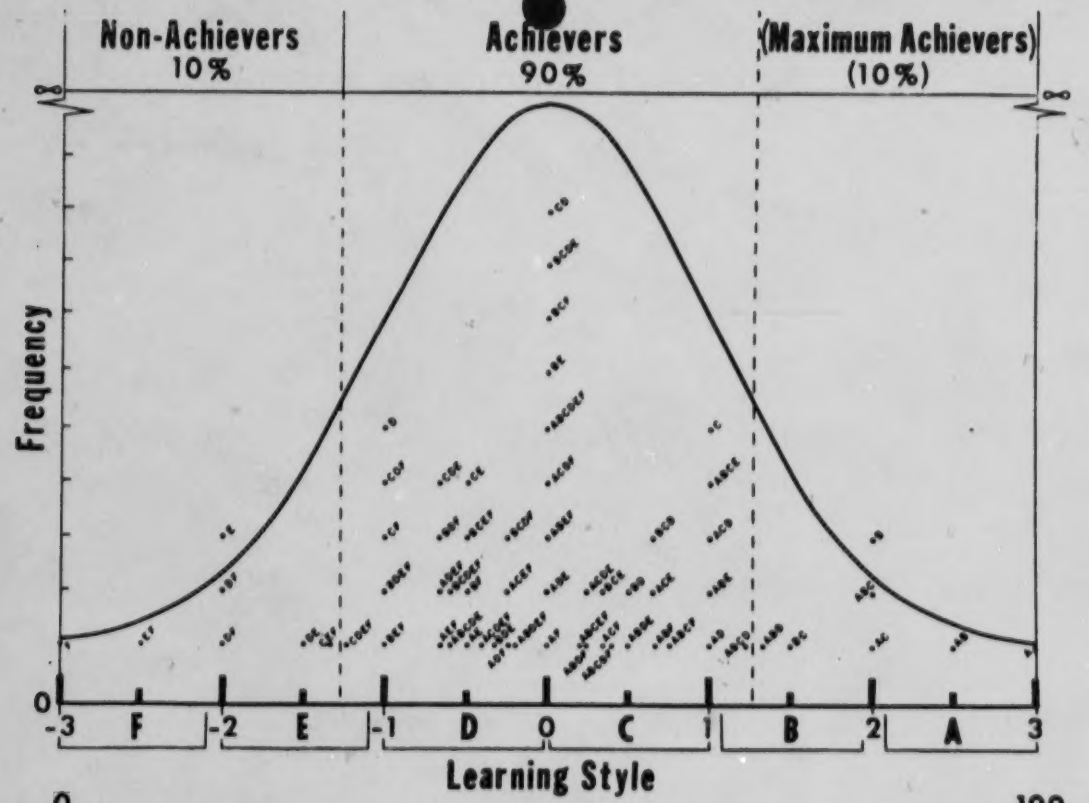
Clearly, then, in order to elicit maximum learning in all its aspects, the program must be written for the learning style of the individual. When this is not done, the percentages in the achiever group shrinks from 90 to some lower figure, say 10 percent. This means that the real question at issue is not "What about the other 10 percent?" but "Why only Style A; why only 10 percent maximum proficiency?"

Seemingly, all that remains to be done is to discover the composition of learning styles and then construct programs to fit. But if, indeed, there is a multiplicity of factors contributing to an individual's learning style, and if there is no reasonably finite limit to the possible combinations of these factors, the result is a unique style for each individual, and this would call for programs by prescription. As a logical conclusion to the tutorial approach, such a condition may eventually obtain through the use of intrinsically programmed electronic teaching machines, but writers of paper-and-pencil programs may have to be content to produce programs using a smaller, more manageable number of instructional styles which favor a range of related learning styles.

Programmers need not mark time waiting for factor analyses to be completed and style limits drawn. Instead, they may start fractionating the target audience by comparing individual results on a program with those on recognized diagnostic instruments of choice, then writing alternate forms of the same program for sub-audiences discovered. In this scheme, a program in mathematics presently aimed at the sixth grade might be broken out in one form as "mathematics for sixth grade pupils who like to

solve puzzles," and in another form as "mathematics for sixth grade pupils who like to draw diagrams," and so on to fit, or nearly fit, the learning style of each child placed in the sixth grade on the usual placement criteria. Admittedly, this is a hit-and-miss approach which serves only to subdivide a target group. Nevertheless, it should effect some improvement in program completion times, achievement scores, and retention times and percentages on the basis that it takes into consideration, albeit crudely, the individual predilection to react in one way as opposed to another. Programmers may not be able, for example, to distinguish Style B from AC or ABC, but as a start, they may be able to distinguish this "phenotype" from the E, DEF, EF, and F complex without necessarily being fully aware of why this is possible.

Thus it may be concluded that until programmers take into account the unique learning style of the individual, they will not have broken the lockstep at all, merely increased, and laudably so, the likelihood that a larger percentage of a given class will be in the achiever group, leaving unanswered the question, "What about the other 10 percent?"



% Content	0	100
Completion Time	∞	0
Retention Time	0	∞
% Retention	0	100

Figure I. Expected performance according to learning style when instructional style favors style A and discriminates most against style F.

AN AIR FORCE JOB SATISFACTION RESEARCH PROJECT THROUGH 1976

R. Bruce Gould
Occupational and Manpower Research Division
Air Force Human Resources Laboratory

This paper presents the status of a comprehensive long-term job satisfaction research project. Ultimate goals of the project are full utilization and retention of qualified Air Force enlisted personnel.

Initiation of the research effort was announced at the 1971 conference of the Military Testing Association (MTA) (Gould & Christal, 1971). A 1973 status report to this association (Gould, 1973) presented the evolving job satisfaction research plan. This paper will present an overview of the research project, update previously reported results, and summarize the current status of the project.

The job satisfaction research approach has concentrated on studying the impact of occupational factors on job attitudes, productivity, and career decisions. As the research progressed, goals of the project were established as follows: (a) to define the dimensions of job satisfaction operating in the Air Force work environment; (b) to measure satisfaction levels on these dimensions; (c) to identify problem areas which have the greatest potential for improvement through satisfaction research; (d) to measure the effects which specific changes in job content have on job attitudes; and (e) to implement job reengineering actions and measure their effects on job attitudes, job performance, and eventual reenlistment decisions. Studies concerning each aspect were concurrently conducted although emphasis has been, of necessity, on the initial steps.

To date, basic studies have been accomplished which (1) surveyed the relevant literature and developed a theoretical approach to satisfaction for use in Air Force job satisfaction research projects (Tuttle & Hazel, 1974); (2) defined the dimensions of job satisfaction operating in the Air Force work environment and developed an experimental job attitude inventory to measure satisfaction levels on those dimensions (Tuttle, Gould, & Hazel, 1975); (3) refined the measurement instrument and developed an operational Air Force Occupational Attitude Inventory (Gould & Christal, 1976); (4) developed an extensive data base containing task-level performance data, background information, job attitudes, aptitude, and loss/gains data for indepth within and between specialty studies (Gould, 1973); (5) developed a

methodology for identifying specialties which are most likely to yield equitable returns from indepth study (Gould, 1974); (6) conducted studies of specific factors associated with dissatisfaction in civil engineer, munitions, and aircraft control and warning ladders and findings have resulted in selected specialty changes (Gould, 1976b); (7) determined that job satisfaction is related to career decisions (Tuttle & Hazel, 1974); (8) developed a location of assignment preference survey and determined preferred geographical characteristics (Tuttle, Brockhaus, & Hazel, 1974; Hazel, Stacy, & Burtch, 1975); (9) found that few airmen are assigned to preenlistment preferred work areas even though those receiving such assignments are more satisfied (Hendrix & Ward, 1975); (10) determined that projecting personnel turnover from self reports of job satisfaction and career intentions was feasible and developed a prediction system (Alley & Gould, 1975); (11) completed a causal analysis of relationships between performance and satisfaction in eight airmen specialties (Hickerson, Hazel, & Ward, 1975); (12) developed a method for determining desirable task experiences of first-line supervisors and impact of experience inadequacies on supervisors and subordinates (Stacy & Hazel, 1975); (13) evaluated the relationship between biographical factors and leisure activities to job attitudes (Burtch & Hazel, 1975); (14) studied longitudinal changes in job attitudes (Gould, 1976a); (15) determined that there were no apparent within career ladder racial differences in airmen work assignments and resulting job attitudes (Christal, 1972); (16) performed a literature review of leadership approaches and developed a theoretical framework for future research into effects of leadership styles on job satisfaction and performance (Hendrix, 1976); and (17) reviewed the job enrichment literature and developed guidelines for future Air Force job reengineering actions (Watson & Zumbro, 1976). Specific comments will be made on major studies, particularly those completed since 1973. Proceedings of this conference will contain the references for detailed treatments of each topic.

Literature Review and Development of a Theoretical Job Satisfaction Model

Tuttle and Hazel (1974) reviewed 400 civilian and 284 military publications and summarized that: (a) job satisfaction is related to tenure, absenteeism, and sick calls but the relationship between satisfaction and performance is inconclusive; (b) job satisfaction is multidimensional, composed of continuous attributes rather than separate satisfaction/dissatisfaction attributes as proposed by Herzberg's (1966) two-factor theory; (c) operationally useful results

are more likely if specific satisfaction dimensions rather than global measures are studied; and (d) no satisfactory theoretical framework or attitude measurement devices were found for use in the military environment. Tuttle and Hazel (1974) developed a theoretical model for use in the Air Force Human Resources Laboratory job satisfaction research.

Development of an Air Force Occupational Attitude Inventory OAI

Tuttle, Gould, and Hazel (1975) identified 35 facets of satisfaction, wrote 348 items to measure the facets in a military work environment, and had 3,000 airmen rate the items on an 8-point rating scale ranging from Extremely Dissatisfied to Extremely Satisfied. Most of the original 35 hypothesized facets were developed from a review of civilian studies. A principal components factor analysis with varimax rotation of the responses found that, in fact, only 15 of the 35 hypothesized dimensions were substantially replicated. Summarizing the differences between the hypothesized and empirically derived job satisfaction factors, work dimensions such as Achievement, Importance, Interest, Job Design, Responsibility, Job Change, Utilization of Abilities, and Variety were all represented by one new factor, the Work Itself. Similarly, items written for facets Supervision Received - Technical and Supervision Received - Human Relations combined to form a Supervisor's Competence factor. While items of aforementioned dimensions combined to form single factors, items in a hypothesized Air Force and Unit Policies and Practices facet divided into five separate factors; Pay and Benefits items formed four factors; and Social Status items formed two. Five new factors were found which took items from multiple sources although the primary source was identifiable. Results suggest that the military work environment and facets relating to job satisfaction differ substantially from their civilian counterparts and thus data collection instruments should be tailored for the military setting rather than using civilian instruments. Additional analyses of the data resulted in reduction of the inventory to 200 items which still measure the domain of variance associated with the full 348 item set (Gould & Christal, 1976). The 8-point scale was changed to a 9-point scale with a midpoint because of skewed rectangular item response distributions. Administration of the 9-point scale and 200 items to 8,000 cases validated the item reduction process and factor analyses (Gould, 1976b).

Relation of Participation in Sports and Leisure Activities to Job Attitudes

The initial version of the OAI contained 74 background items and for the 3,000 case sample, stepwise regression techniques were used to identify items which had unique relationships with job attitudes. Several items such as race; age when obtained first regular paying job; parental mobility, education, socioeconomic status, or individual's contribution to parent's income; birth order; or draft vulnerability when entered service did not contribute to job attitude prediction when factors such as age, marital status, academic standing, aptitude, and selected leisure activities were held constant. In separate analyses, Burtch and Hazel (1975) analyzed relationships of 18 past and current leisure time activities and found that satisfied personnel tended to have a greater preference for western music and spend more time fishing than dissatisfied personnel who like rock music and play more tennis. Otherwise, no sports and leisure activity differences were found between satisfied and dissatisfied airmen.

Development of An Extensive Data Base

The Air Force Occupational Survey Program has routinely collected job attitude data with each survey administered since 1966. Of 236 Air Force enlisted career ladders, survey data exist for about 250,000 airmen in 200 ladders. The survey data has been matched with personnel files and an extensive single longitudinal data file which contains task-level definitions of jobs performed, job attitude statements, background, loss/gains, aptitude and career progression information. The file is available for within and between specialty analyses, identification of productive specialties for indepth analysis, and Air Force-wide change or trend analyses. Selected studies using the file will be reported; potential for additional comprehensive studies now exists.

Indepth Studies of Specific Career Ladders

A number of specific career ladder job satisfaction studies have been conducted and reported (Gould & Christal, 1971; Gould & Christal, 1973; Gould, 1976b). To summarize their findings, factors related to reported dissatisfaction are essentially unique to each career ladder. There are, however, a few relationships which have been consistently found. Those job incumbents whose jobs are more demanding and varied and are most related to their aptitude, education, and Air Force training are more satisfied with their jobs and tend to express positive predispositions toward Air Force careers. Regression studies indicate that job difficulty, number of tasks performed, and aptitude

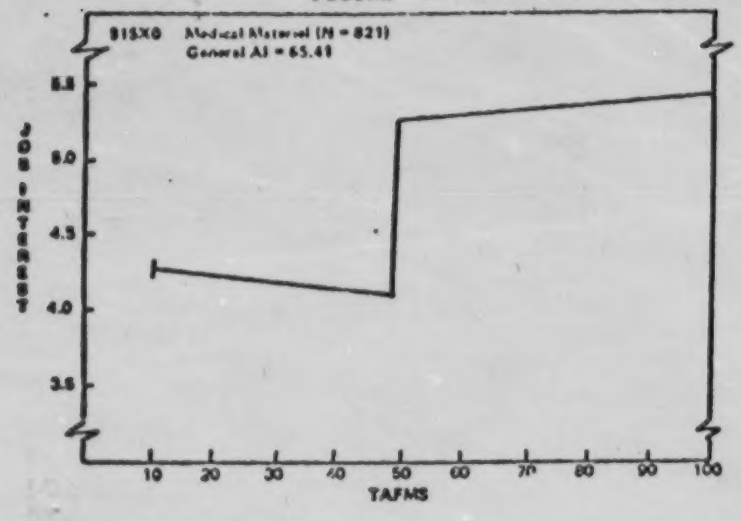
requirements of the job are positively related to job attitude while aptitude, education, and time in service for 1st-term airmen are negatively related. Time in service is positively related to job attitude for career airmen. Most indepth studies have identified problems whose solutions require a combination of job reengineering and management decisions. However, the cost of such studies is high and not all have resulted in recommendations for operational remedies. Thus a procedure for identifying ladders with greatest potential for operational returns from indepth study was developed. Operational returns are defined as improved reenlistments, job performance, or career progression.

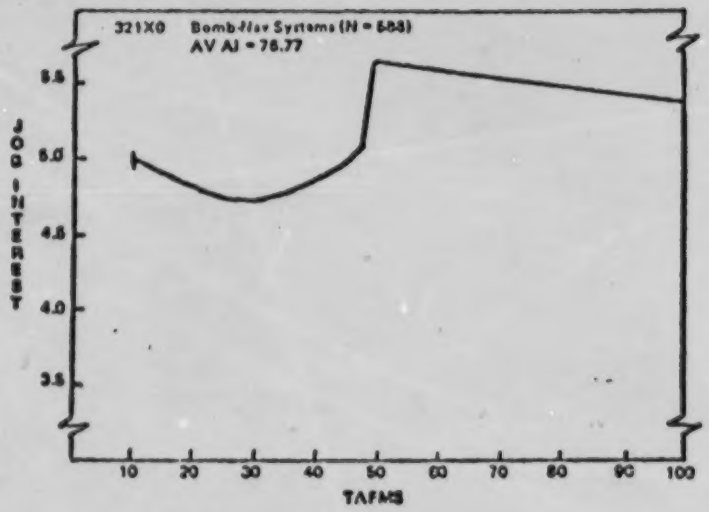
Identification of Specialties with the Greatest Potential for Improvement Through Satisfaction Research

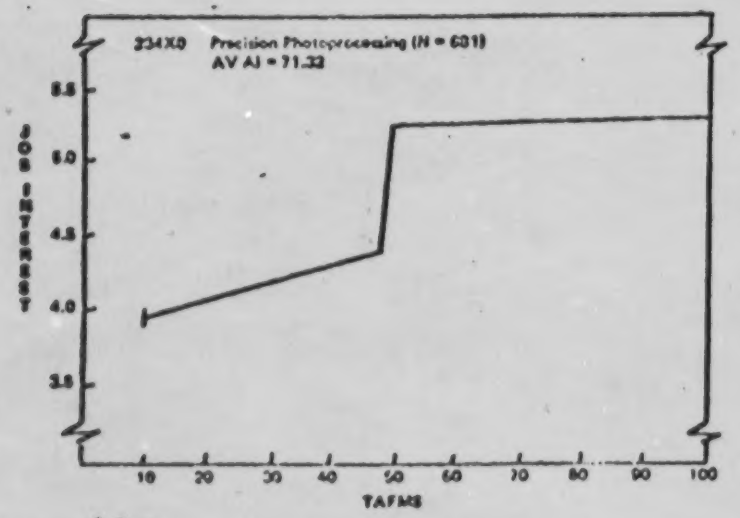
The 1973 (Gould, 1973) paper to this conference presented a regression model which was used to display cross-sectional attitude data and make inferences of job attitude relationships to career decisions and job progression. Figure 1 presents job attitude time in service relationships for four specialties. To interpret the ladder profiles, compare the profile of the Medical Materiel ladder in the upper left with the Operating Room Technicians in the lower right. Job interest decreases for Medical Materiel airmen as they approach the reenlistment decision point at their 48th month. The jump from the 48th to the 49th month suggests that those reporting uninteresting jobs did not reenlist. The magnitude of the jump is interpreted to indicate the degree of relationship between job interest and the reenlistment decision. A ladder exhibiting such a profile has a high likelihood of impact on positive job attitudes and career decisions through job reengineering or other corrective actions to produce a positive, rather than negative, accelaration in attitude as the reenlistment point is approached. Ladders with profiles such as the Dental Technicians show little relationship between interest and reenlistment. The Dental ladder shows no potential return for efforts made to improve the jobs.

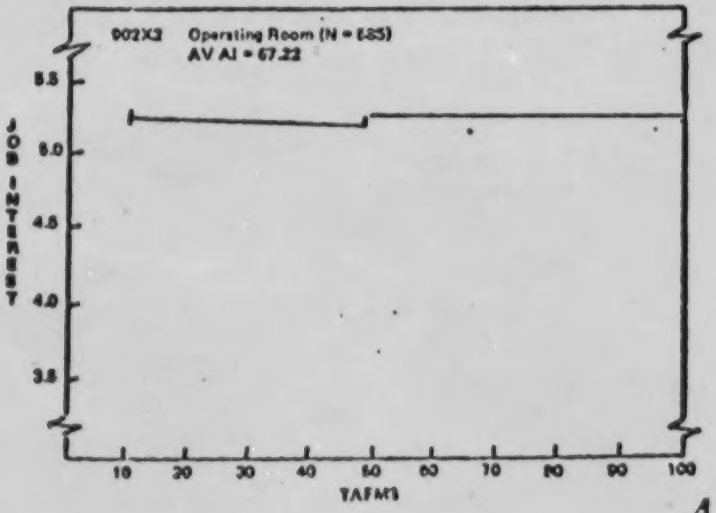
Thus far, cross-sectional data has been interpreted in a longitudinal fashion, a risky procedure at best. Alternate explanations of the data focus on the concept of "cognitive dissonance" (Festinger, 1957) where the attitude change between the 48th and 49th month is believed to result from individuals

FIGURE 1. JOB INTEREST VS TAFMS (HOLDING APTITUDE CONSTANT)









373

423

424

making some past behavior, in this case reenlistment, consistent with the present situation to reduce internal or intrinsic conflict. However, the "cognitive dissonance" interpretation was refuted and the earlier interpretations validated by a longitudinal study of seven career ladders. Longitudinal data demonstrated that the 49th month jump was the result of a "residualization process" where dissatisfied airmen left the service. Those remaining after the reenlistment decision point had more positive job attitudes at all points leading up to the reenlistment decision suggesting also that those airmen who perceived job progression were the airmen most likely to reenlist.

Forecasting Reenlistment Decisions

Another study into attitude/reenlistment relationships (Alley & Gould, 1975), traced 53,000 first-term airmen respondents who had completed job attitude and career decision statements at various years of service prior to reaching the actual reenlistment decision point and determined their reenlistment decisions. Purpose of the study was to evaluate the feasibility of predicting reenlistments by career ladder. A model seeking procedure developed a model for predicting a dichotomous 1 = reenlisted and 0 = was discharged criterion. The final model took into account the time interval between survey administration and time of decision and the career intent statement. Predictive validities across the first four years for the total sample were .14, .17, .37, and .53 respectively. Actual functional relationships between career intent responses and career decisions appeared to be curvilinear. When the model was used to project reenlistments for the 30 most populous career ladders and the projections compared with actual reenlistments, errors of prediction of from 9.4 percent for airmen in their first year to only 2.1 percent for airmen in their fourth year of service were obtained. Using job attitude statements, multiple R's for predicting the reenlistment/discharge criterion for the 30 ladders ranged from .12 to .31 using job interest as the predictor and from .11 to .30 using perceived use of talents and training as the predictor. However, the job attitude statements did not add significantly to the prediction system in the presence of career intent statements (Gould, 1976a). This finding demonstrates the considerable relationship between job attitude and career decisions and that in the absence of career intent statements, job attitude statements can be useful in forecasting reenlistment rates.

Preferred [illegible] Assignment Effect on Job Satisfaction

Hendrix and Ward (1975) studied the relationship between preenlistment job preferences and actual work assignment for 20,000 airmen. They found that only 15 percent of the enlistees from 1966 to 1974 received any of their three preferences, but reports of job interest and perceived talents and training utilization were significantly more positive for airmen receiving one of their preferences. The job attitude statements were made from six months to four years after the preference statements.

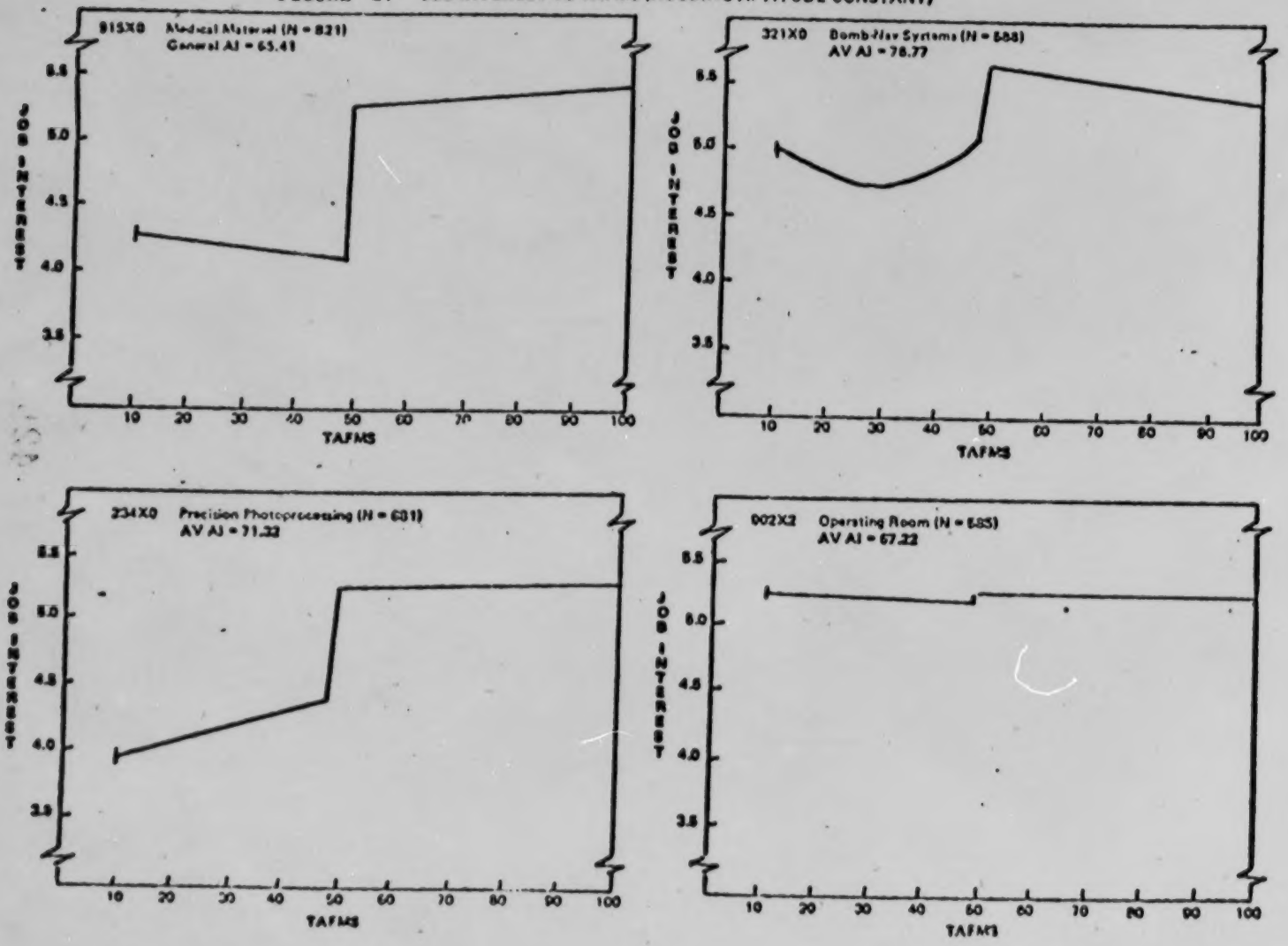
Causal Relationships Between Job Attitude and Performance

A longitudinal study of 1,352 airmen in eight specialties investigated the relationships between several measures of job satisfaction and job performance (Hickerson, Hazel, & Ward, 1975). Several complex analyses concluded that the relationship between satisfaction/performance is extremely complex and tends to be career ladders specific. In general, a circular causality relationship seems to exist where satisfaction and performance influence each other.

Leadership Styles and Effects on Job Satisfaction and Performance

A study by Stacy and Hazel (1975) developed procedures for identifying tasks which supervisors must have experience in performing in order to adequately supervise their performance. Effects of non-experience on both the incumbent's and the supervisor's job attitude were evaluated. In addition to the technical competance of the supervisor, his leadership style is presumed to influence his effectiveness as a manager, where effectiveness is measured in terms of incumbent performance and satisfaction. Hendrix (1976) conducted an extensive literature review and developed an eclectic model titled "The Three-Component Leadership Effectiveness Model." The model considers leadership effectiveness to be an interactive function of the criterion selected, the leadership style employed, and the situational environment which includes the leader's subordinates, and other personnel in the environment. The strength of the model is that: (a) it provides a useful descriptive framework for depicting the phenomenon of leadership or management behavior; (b) it presents leadership as a dynamic decision-making process which requires modification as the environment changes; and (c) it lends itself to testing and therefore can be verified. Current projects are now validating the model.

FIGURE 1. JOB INTEREST VS TAFMS (HOLDING APTITUDE CONSTANT)



373

427

428

Preferred Job Assignment Effect on Job Satisfaction

Hendrix and Ward (1975) studied the relationship between preenlistment job preferences and actual work assignment for 20,000 airmen. They found that only 15 percent of the enlistees from 1966 to 1974 received any of their three preferences, but reports of job interest and perceived talents and training utilization were significantly more positive for airmen receiving one of their preferences. The job attitude statements were made from six months to four years after the preference statements.

Causal Relationships Between Job Attitude and Performance

A longitudinal study of 1,352 airmen in eight specialties investigated the relationships between several measures of job satisfaction and job performance (Hickerson, Hazel, & Ward, 1975). Several complex analyses concluded that the relationship between satisfaction/performance is extremely complex and tends to be career ladders specific. In general, a circular causality relationship seems to exist where satisfaction and performance influence each other.

Leadership Styles and Effects on Job Satisfaction and Performance

A study by Stacy and Hazel (1975) developed procedures for identifying tasks which supervisors must have experience in performing in order to adequately supervise their performance. Effects of non-experience on both the incumbent's and the supervisor's job attitude were evaluated. In addition to the technical competance of the supervisor, his leadership style is presumed to influence his effectiveness as a manager, where effectiveness is measured in terms of incumbent performance and satisfaction. Hendrix (1976) conducted an extensive literature review and developed an eclectic model titled "The Three-Component Leadership Effectiveness Model." The model considers leadership effectiveness to be an interactive function of the criterion selected, the leadership style employed, and the situational environment which includes the leader's subordinates, and other personnel in the environment. The strength of the model is that: (a) it provides a useful descriptive framework for depicting the phenomenon of leadership or management behavior; (b) it presents leadership as a dynamic decision-making process which requires modification as the environment changes; and (c) it lends itself to testing and therefore can be verified. Current projects are now validating the model.

Current Studies

Current studies are (1) analyzing OAI returns of 7,800 airmen for between and within career ladder satisfaction differences and identification of ladder specific job factors which impact on career decisions; (2) identifying additional career ladders for indepth research where manipulable job content factors are related to unacceptably low reenlistment rates; (3) expanding the methodology for identifying critical task experiences for supervisors; (4) in two specialties, evaluating changes in job content which have occurred normally in individuals' jobs through job progression or reassignment during a 13 to 18 month time period - job changes will be related to job attitude changes during the same period; (5) determining the amount of variance in job attitudes and performance ratings which is related to motivational and interest characteristics of individuals and is not a function of the work environment and therefore not subject to change through work interventions; (6) elaborating on the relationships between work environment and job attitudes and productivity; and (7) developing techniques for first-line supervisors to increase the intrinsic motivation and thus productivity and satisfaction of subordinates.

Conclusion

The job satisfaction research project has been highly productive. The prognosis for the projected goal of full utilization of personnel, retention of qualified airmen, and maintenance of critical skills is promising. From a cost effective standpoint, and in an era of increasing personnel costs and decreasing budgets and strength authorizations, small percentage gains in productivity and retention of trained personnel will produce substantial returns from minimal research efforts. Later years should see additional applications of the evolving technology.

REFERENCES

Alley, W.E. & Gould R.B. The feasibility of estimating personnel turnover from survey data - A longitudinal study. AFHRL-TR-75-54, AD-A018 777. Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, October 1975.

Burtch, L.D. & Hazel, J.T. The relation of sports and leisure activities participation to airmen job attitudes. AFHRL-TR-75-66, AD-A020 091. Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, November 1975.

Christal, R.E. Analysis of racial differences in terms of work assignments, job interest, and felt utilization of talents and training. AFHRL-TR-72-1, AD-741 758. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, January 1972.

Festinger, L. A theory of cognitive dissonance. New York: Harper & Row, 1957.

Gould, R.B. Review of Air Force job satisfaction and career development research. Proceedings, 15th Annual Conference of the Military Testing Association, U.S. Air Force, San Antonio, October 1973, pp 560-572.

Gould, R.B. Review of Air Force job satisfaction research. Paper presented at American Psychological Association Convention. New Orleans, August 1974. (Available from ERIC, Accession number ED 099 753).

Gould, R.B. Longitudinal inferences of job attitude/tenure relationships from cross-sectional data. AFHRL-TR-76-46, AD- . Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, July 1976a.

Gould, R.B. Review of an Air Force job satisfaction research project: Status report through September 1976. AFHRL-TR-76-75. Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, 1976 (in press).

Gould, R.B. & Christal, R.E. Enlisted job satisfaction in the Air Force: A study at the task level. Proceedings of the 13th Annual Conference, Military Testing Association, U.S. Marine Corps, Washington, D.C., September 1971, Pp. 323-333.

Gould, R.B. & Christal, R.E. VARSEL: Variable selection for multiple purpose prediction systems in the absence of external criteria. AFHRL-TR-76-6, AD- . Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, May 1976.

Hazel, J.T., Stacy, W.J., & Burtch, L.D. Assignment location preferences of airmen. AFHRL-TR-75-50, AD-A017 167. Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, October 1975.

Hendrix, W.H. Contingency approaches to leadership: A review and synthesis. AFHRL-TR-76-17, AD- . Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1976.

Hendrix, W.H. & Ward, J.H., Jr. Preferred job assignment effect on job satisfaction. AFHRL-TR-75-77, AD- . Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, December 1975.

Herzberg, F. Work and the nature of man. New York: The World Publishing Company, 1966.

Hickerson, K.A., Hazel, J.T., & Ward, J.H., Jr. A causal analysis of relationships between performance and satisfaction in eight airman specialties. AFHRL-TR-75-57, AD-A020 542. Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, October 1975.

Stacy, W.J. & Hazel, J.T. A method of determining desirable task experiences for first-line supervisors. AFHRL-TR-75-23, AD-A017 471. Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, August 1975.

Tuttle, T.C., Gould, R.B., & Hazel, J.T. Dimensions of job satisfaction: Initial development of an AF occupational attitude inventory. AFHRL-TR-75-1, AD-A014 796. Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1975.

Tuttle, T.C. & Hazel, J.T. Review and implications of job satisfaction and work motivation theories for Air Force research. AFHRL-TR-73-56, AD-782 443. Lackland AFB, Tex.: Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

Tuttle, T.C., Brockhaus, W., & Hazel, J.T. Development and feasibility test of a method to study location assignment preferences of airmen. AFHRL-TR-74-40, AD-781 041. Lackland AFB, Tex.: Occupational Research Division, Air Force Human Resources Laboratory, April 1974.

Watson, T.W. & Zumbro, P.A. An evaluation of job enrichment with implications for Air Force job redesign research, and An annotated bibliography of job enrichment and related literature. AFHRL-TR-76- , AD- . Lackland AFB, Tex.: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, 1976 (in press).

A COMPUTER ASSISTED TEST CONSTRUCTION SYSTEM

by

Vergil M. McIntosh
Educational Specialist

Extension Course Institute
Air University

Preparing examinations includes much clerical work and typing which can be automated. I am sure this is an obvious statement to you. To assist in this work a number of users have already developed Computer Assisted Test Construction Systems which they claim are cost effective and improve test quality.

Several years ago, the Air Force Extension Course Institute began investigating possible applications of computers in test construction work. Our situation is quite different from most users because we have many courses and examinations in use at any one time and our item banks are relatively small. We have almost 400 different courses comprising approximately 1250 volumes of course material. We have over 3000 volume review exercises and course examinations in the field. Item banks average about 200 items each. Keeping these characteristics in mind we undertook to develop our CATC system to assist in this great volume of work.

I shall, in the next few minutes, give a brief description of the major programs or software packages and the purposes and output of each. We are just beginning to use the system but the programs are all operational with the exception of the item selection program which we still want to revise. We operate from a time-sharing system using the Honeywell 6000. Our remote terminals are a GE TermiNet 300 and an IBM Mag Card Selectric.

SYSTEM FILES

Item Bank

Three input files are required for our CATC system. The first is an item bank for each volume of a course. Figure 1 shows a section of an item bank. Each item is coded to show the (1) learning outcome, (2) sequence number under each learning outcome, (3) the revision number of the item, (4) the option print order if the options cannot be randomized, and (5) page reference. The first three codes are used as line numbers in the item bank. This provides a unique number for each item and automatically places it at the correct position in the item bank.

Item Data

The second file is an item data file which includes coded information about each item. Figure 2 is a section of an item data file. This data includes:

- learning outcome number
- sequence number under the learning outcome
- revision number
- taxonomic level
- item importance
- item use - suitable for a Volume Review Exercise, a Course Examination, or either
- estimated item difficulty
- estimated item discrimination
- item efficiency index
- item reading difficulty level, and
- records of examinations on which the item was used

Selected Items

The third file is a list of the items to be used on the examination and will be described later.

The following program packages then use these files to print out a camera-ready copy of the examination.

SYSTEM PROGRAMS

Item Catalog

The first program prints out a catalog of the items in an item bank along with the associated data, as shown in Figure 3. The test constructors use this to determine the need for new items or for item revisions. The appropriate changes are then made in the item bank. The Editor System is used if the changes are minor such as changing words or punctuation. Major changes are usually made by inputting the appropriate line number and the new line or lines. The data listed at the top of each item in the item catalog is the same as that I mentioned in describing the item data file.

Once the item banks are edited we are ready to select the items to use.

Item Selection

The test constructor has the option of either making item selections from the item catalog, or having the computer make the selections.

If he makes the selections, he lists the code number of the items, which are then entered on disk file. The print program then uses this file to match with items in the bank to select the items to print.

If he elects to have the computer make the selections, he sets the parameters by responding to a series of questions at the beginning of the program. The parameters are:

- the beginning and ending learning outcome number for each item bank
- the number of items desired on the examination
- the minimum and maximum number of items per learning outcome
- the minimum and maximum test difficulty
- minimum percent of application items
- minimum percent of understanding items
- maximum duplication permissible with items which were used on the Volume Review Exercise
- maximum duplication with the parallel course examination
- minimum item importance
- minimum item efficiency

The computer scans the item data file, small groups of items at a time, until it finds the required number of items that fall within these parameters. If it fails to find enough items after four searches, some of the parameters are relaxed. If it fails to find enough items after eight searches, it gives up and goes on to the next group of items. The program also creates a file of the items to be used on the examination.

Because of the large number of parameters we are using and the relatively small item banks, we have not yet finalized this program. We need to either add more items to some item banks or adjust the program to bypass more parameters.

Assign Items and Print Statistical Report

The Extension Course Institute prepares a Volume Review Exercise, an open-book examination over each volume of a course. We also prepare two forms of a closed-book examination over each course, these we call Course Examinations. It is important that the Course Examinations be parallel forms. This next program is used only for course examinations. It assigns items to the two forms so they will be of equal difficulty. It also computes and prints a statistical report as shown in Figure 4. The statistics reported are:

- average item difficulty
- average item discrimination
- mean score
- standard deviation
- test reliability
- failure rate

If the statistics appear satisfactory, we proceed with the next step, which is to print the camera-ready copy of the examination.

Printing

Once the file of items to be used is created, we proceed to print the camera-ready copy. This is printed on the IBM Selectric terminal because of the terminal's very good printing quality. The program locates the items in the item bank and prints them in the proper format. Figure 5 shows the printout format. This program performs the following operations.

- locates the selected items in the item bank
- prints the items in the proper format
- marks the top and bottom of each page
- numbers each item
- determines when to terminate a page and numbers each page
- randomizes the order in which the options are printed
- offers the test constructor the option of specifying the order to print the distractors

- numbers the item and leaves the proper space to enter the item if figures or special characters are required

- creates a file of information to print the answer key

The remote terminal is a relatively slow means of printing; however, it does print in upper and lower case and it has much better print quality than a line printer. It will print a 100 item examination in about 40 minutes compared to about two days for a typist to type the examination.

Print Answer Key

The answer key is printed next using the file prepared by the print program. Figure 6 is a section of this printout. For each item it shows the (1) item number, (2) correct response, (3) learning outcome and, (4) page reference. Printing an answer key requires less than five minutes as compared to a couple of hours to manually prepare a key.

The answer key is presently keypunched and sent to our scoring section where answer sheets are scored by batch processing. To eliminate this keypunch operation, we will have the scoring unit read our disk file to obtain the score key information.

Other Programs

This completes the description of the major programs in the CATC system. There are other minor programs, such as programs for file maintenance or to check data files.

CONCLUSIONS

Evaluation

The system operates as designed, although we have not used it long enough to realize many of the benefits. It takes some time for people unaccustomed to a computer to accept it. This is a finding in all CATC systems. We feel that the system will effectively reduce manpwoer requirements and improve test quality, and free test constructors from routine clerical tasks.

Observations

- About the same length of time is required for a typist to enter items in disk storage as it requires for normal typing. Since the items are used several times, this results in a significant reduction in typing.

- A remote terminal will print a 100-item examination in thirty to forty-five minutes as compared to about two days for a typist to type an examination.

- After the item banks are properly edited, the examination is printed error free.

- Error free answer keys are produced, whereas errors occassionally crop in when answer keys are constructed manually.

- A neat copy is produced which is suitable for camera-ready reproduction.

- Much better control is provided over duplication with VREs and other forms of the CE.

- The statistical report gives the test constructor a reasonably accurate prediction of how the test will perform.

- Updating item files and making corrections is a relatively simply operation.

- More frequent revisions of examinations will become feasible. This should reduce the likelihood of test compromise, and also result in fewer deleted items because of changes in the text materials.

- The item selection program should greatly enhance the selection of appropriate items according to the desired parameters.

Problems

The major problem is file space for item banks. We presently input the banks on magentic disk storage, and when the examinations have been completed we run the banks on cassette tapes for permanent storage. The most desirable method would be to have enough space to keep the banks on disks. We have found that a typical bank of about 200 items requires about 40 to 50 blocks of storage, so this amounts to a great storage requirement for the number of examinations that we have. Some of the alternatives we are considering are remotes with diskette or "floppy disk" storage, removable disk packs, or possibly magnetic tape at the computer site.

Probably the next most important problem is for test constructors and typists to become familiar with the system and to gain confidence in it. Very few of our test constructors are familiar with computer operation and therefore lack the necessary confidence.

Some typists are apprehensive about learning the system and view it as something mysterious. There is also some concern that the "machine" will take over their jobs.

We have some down-time with time-sharing; however, the system seems to be improving. Although this is sometimes an aggravating problem, it is not too serious.

```
2001100 2:3:2134:1
2001101 Which regulation establishes training requirements for egress
2001102 personnel?
2001103 a.: AFR 66-51.
2001104 b.: AFR 66-1.
2001105 c.: AFR 127-100.
2001106 d.: AFR 127-101.
2001107 XXXX
2002100 1:5:1:1
2002101 Who must certify your egress systems training?
2002102 a.: The Chief of Maintenance.
2002103 b.: Your supervisor.
2002104 c.: The training control section.
2002105 d.: The field training detachment commander.
2002106 XXXX
2003100 2:4:1:1
2003101 Where is AF Form 1098, Special Task Certification and
2003102 Recurring Training, kept?
2003103 a.: Duty section.
2003104 b.: Orderly room.
2003105 c.: Maintenance control.
2003106 d.: Training office.
2003107 XXXX
```

Note: Legend for top line--200=learning outcome nr; 1=sequence nr; 1=revision nr; 00=line sequence; 2=print format; 3=nr lines to print item; 2134=distractor print order --a "1" would indicate distractors can be randomized; 1=page reference. XXXXs are flags to signal end of item.

Figure 1. Item Bank.

	VOL	LO	SEQ	REV	TAX	IMP	EX	EEZ	EDS	%EF	IRD	VRE	CE#1	CE#2	PAG
V C C	21	200	1	1	1	2	1	70	50	55	.0	25.001	.000	.000	1
V C C	21	200	2	1	1	1	3	86	28	66	.0	.000	.000	.000	1
V C C	21	200	3	1	1	1	3	84	26	54	.0	.000	.000	.000	1
V C C	21	200	4	1	1	2	3	88	24	66	.0	25.002	.000	.000	1
V C C	21	200	5	1	2	2	3	90	18	60	.0	25.003	.000	.000	1
V C C	21	200	6	1	2	2	3	76	42	58	.0	25.004	.000	.000	1
V C C	21	201	1	1	1	2	3	80	32	53	.0	25.005	.000	.000	2
V C C	21	201	2	1	2	2	3	78	34	51	.0	.000	.000	.000	2
V C C	21	201	3	1	2	3	3	82	2[illegible]	51	.0	25.0[illegible]	.000	.000	2

Note: See Figure 3 for explanation of headings.

Figure 2. Item Data File.

ITEM BANK FOR COURSE 42252
09/16/76

VOL	LO	SEQ	REV	TAX	IMP	EX	EEZ	EDS	%EF	IRD	VRE	CE#1	CE#2	PAG
2	200	1	1	1	2	1	70	50	55	.0	25.001	.000	.000	1

Which regulation establishes training requirements for egress personnel?

a.AFR 66-51.
b.AFR 66-1.
c.AFR 127-100.
d.AFR 127-101.

VOL	LO	SEQ	REV	TAX	IMP	EX	EEZ	EDS	%EF	IRD	VRE	CE#1	CE#2	PAG
2	200	2	1	1	1	3	86	28	66	.0	.000	.000	.000	1

Who must certify your egress systems training?

a.The Chief of Maintenance.
b.Your supervisor.
c.The training control section.
d.The field training detachment commander.

Note: VOL=volume; LO=learning outcome; SEQ=sequence nr; REV=revision nr; TAX=taxonomic level; IMP=item importance; EX=item use-VRE, CE or either; EEZ=item difficulty; EDX= estimated item discrimination; %EFF=item efficiency; IRD= item reading level; VRE=form nr of VRE and item nr; CE=forms of CEs on which item was used; PAG=page reference.

Figure 3. Item Catalog.

COURSE EXAMINATION STATISTICAL ESTIMATES
COURSE 42252 FORM 33 FORM 34
09/16/76

NR ITEMS=,	124	124
AVG EASE =	.795	.794
AVG ITEM DISC=	.322	.324
MEAN OF RAW SCORES =	98.63	98.49
STANDARD DEVIATION =	18.81	18.91
PASS/FAIL POINT =	74.4	74.4
EST FAILURE RATE=	.09	.09
EST RELIABILITY =	.951	.951

Figure 4. Statistical Report.

MULTIPLE CHOICE

Note to Student. Consider all choices carefully and select the best answer to each question.

1. (200) Which regulation establishes training requirements for egress personnel?

   a. AFR 66-1.
   b. AFR 66-51.
   c. AFR 127-100.
   d. AFR 127-101.

2. (200) For the egress system you maintain and inspect, recertification training is required every

   a. 3 months.
   b. 6 months.
   c. 12 months.
   d. 18 months.

3. (200) Various maintenance personnel require access to aircraft cockpits containing egress systems. What type of egress training must all such maintenance personnel receive?

   a. Recertification training.
   b. Advanced egress system training.
   c. Basic egress system training.
   d. Egress system familiarization safety training.

Figure 5. Camera-Ready Copy of Examination.

ECI VRE ANSWER KEY

COURSE NR	VOLUME NR	VRE/CE FORM	LAST ITEM NR	DATE OF ANSWER KEY
42252	2	25	100	09/16/76

NR	ANS	GUIDE OR LO	REF	NR	ANS	GUIDE OR LO	REF	NR	ANS	GUIDE OR LO	REF	NR	ANS	GUIDE OR LO	REF
1	2	200	1	2	3	200	1	3	4	200	1	4	2	200	1
5	1	201	2	6	4	201	2	7	4	201	5	8	2	201	5

Note: NR=item number; ANS=correct response; GUIDE OR LO= reference guide nr or learning outcome; REF=page reference.

Figure 6. Answer Key.

DYNA: Dynamic Programming Analysis and the Distribution of Effort

by

William J. Moonan, Ph.D.
Consultant for Statistics

Navy Personnel Research and Development Center
San Diego, California 92152

A Presentation Made To

The 1976 Military Testing Association Conference
Pensacola, Florida

October 1976

CONTENTS

1.	INTRODUCTION	1
2.	MULTI-STAGE DECISION PROCESSES	1
2.1	A Christian MSDP	3
3.	DYNAMIC PROGRAMMING	4
3.1	Recursion Relations	4
3.1.1	Backward and Forward Recursion	4
3.2	Optimality Principle	4
3.3	Transition	5
3.4	Formulation of Dynamic Programming Analysis	5
3.5	Methods of Solution	6
3.5.1	Analytic Solutions	6
3.5.2	Tabular Solutions	6
3.6	The Dynamic Programmers	6
3.6.1	The Theorists	6
3.6.2	The Literature	8
3.7	Some Applications	9
4.	ALLOCATION PROBLEMS	10
4.1	The General Allocation Problem	10
4.2	The Distribution of Effort Problem	10
4.3	The Curriculum Problem	10
4.4	Methods of Reward Determination	13
4.4.1	Token Method	13
4.4.2	Probability Ratio Method	13
4.4.3	Other Methods	13
5.	SOLVING PAGE'S CURRICULUM PROBLEM	14
5.1	The Problem	14
5.2	The Data	14
5.3	Multi-Stage Decision Process Aspects of the Curriculum Problem	16
5.4	Dynamic Programming Analysis Aspects of the Curriculum Problem	16
5.5	A Solution Procedure	16
5.5.1	The Input Phase	19
5.5.2	The Process Phase	19
5.5.3	The Output Phase	21
6.	SUMMARY	29
7.	REFERENCES	30

1. INTRODUCTION

There exist, in everyday life, in the fields of personnel research, personnel management, business, education, engineering, industry and in many other endeavors, a class of situations which we may call "multi-stage decision processes" (MSDP). These MSDP are distinguishable, more or less, by the sequential nature of the order of operations of the process (hence are operational dynamic rather than time dynamic). A few examples are: the investment of capital, the determination of routes, personnel advancement and distribution, the design of facilities and reactors, the production of materials, the extraction of minerals from ores, the control of space vehicles and the allocation of resources, in general. These MSDP are usually important to their parent endeavor since their effective management may be vital to the successful operation of an enterprise. Hence it is important to efficiently control or even optimize operations in a MSDP. A mathematical and computational tool for accomplishing this was developed by Professor Richard Bellman in 1952 and is known as Dynamic Programming (DP).

The purpose of this report is to:

a. Describe MSDP and DP in general and specific terms

b. Indicate the major contributors and contributions to the subjects and to cite relevant literature.

c. Describe a particular allocation problem type (the distribution of effort), relate it to a personnel problem and show how this problem is solved by Dynamic Programming.

2. MULTI-STAGE DECISION PROCESSES

During and after World War II it was realized that a large number of different types of activities were fundamentally similar in that their operation required the specification of a sequence of decisions. Examples of this phenomenon occur in such activities as inventory, production, allocation, scheduling, design and baseball. The sequence of decisions is described by the term "multi-stage decision process."

To determine the sequence of decisions, called a policy, for such activities, a new and important mathematical theory, called Dynamic Programming, was created by Professor Richard Bellman while at the RAND Corporation in Santa Monica, California. So fundamental to this development is the notion of a multi-stage decision process that it is difficult to separate the two ideas.

Activities which may be characterized as a MSDP do not simply appear with the MSDP label attached. Sometimes it is difficult to identify activities or situations as being members of this class. Bellman has been particularly resourceful and clever in doing this, even to showing

(1962) that the old, familiar puzzle of three cannibals and three missionaries crossing a river with the aid of a two-seated boat, is a problem of this class and is solved by DP.

Experience at using DP to determine MSDP policies is probably the best way of becoming familiar with MSDP. To get started perhaps the following descriptions will be helpful.

MSDP: A multi-stage decision process is any sequentially controllable situation recognized as, or part of, a system, procedure, process, activity, operation or problem which may be characterized by the following aspects:

STATES: Are the information that completely specifies the instantaneous condition of the process. The information conveniently expressed and symbolized by mathematical variables, X, perhaps vector-valued, and available at certain sequence points, called stages, of the process.

STAGES: Are those points, in the sequence of the process, at which it is possible to make a decision which controls the process enabling it to change state, earn a reward or suffer a loss.

DECISIONS: Are those actions or variables of choice which affect the process as reflected through changes in the state variables and with which are associated either positive, negative or no reward.

REWARDS: Are those measures of process performance generated over one stage of the process and depend only on the starting and ending states for that stage and therefore may be added for a sequence of stages.

For some problems, rewards are replaced by costs of operation of the process and the type of optimization used is minimization.

CONSTRAINTS: Are those restrictions, often expressed as inequalities, placed on state or decision variables, or their functions, that insure that decisions for the process will be determined under desired conditions. The restrictions define a range of admissible values of the state and decision variables called the state space and the decision space.

OPTIMIZATION: MSDP are analyzed for the purpose of determining that sequence of decisions (called a policy) that will optimize (either minimize or maximize) the total reward for the entire MSDP.

A MSDP is a process with many stages which may either be continuous or discrete. We describe the discrete stage case as follows:

At the beginning of stage n a decision maker obtains information about the process which he is operating and makes a decision. The state of the process changes to a given new state at stage n+1. The new state may depend upon (1) the previous state, (2) the decision made at stage n, and perhaps, (3) some stochastic variables. The decision maker's objective is to make a sequence of decisions so as to optimize some criterion (objective) function for the entire process.

2.1 A Christian MSDP

The author begs your indulgence for departing briefly from his usual statistical, mathematical and scientific themes and introducing a metaphysical theme purporting to indicate a MSDP characterized in Jesus' Sermon on the Mount, THE BIBLE (K.J.V., Matthew 5:1-9). Our metaphysical point of view is that the first seven Beatitudes describe a set of seven states of the process of achieving Spiritual health and happiness. We list below these seven states and their corresponding rewards:

Nr.	State	Reward
1	Poor in Spirit	Theirs is the Kingdon of Heaven
2	Mourning	They shall be comforted
3	Meekness	They shall inherit the earth
4	Hungering and thursting after righteousness	They shall be filled
5	Merciful	They shall obtain mercy
6	Pure in heart	They shall see God
7	Peace-making	They shall be called the children of God

Although these states have an order there is no indication in the Biblical text that such a sequence is either necessary or desirable. We merely point out the existence of this unusual illustration and note also that the personal decisions required to enter or leave the states are not explicitly recorded.

3. DYNAMIC PROGRAMMING

DP is formally presented by Bellman (1957, 1961, 1962) for iteratively finding the optimal decisions for multi-stage decision processes. Stated so simply this description belies Bellman's outstanding development. His technique is based on functional equations. Bellman's Principle of Optimality, leading to DP, often uses the computer for solutions but this is not universally necessary.

3.1 Recursion Relations

Are mathematical equations of the form

(1) $f_n = r_n + f_{n-1}$; $n = n, n-1, \ldots, 2.1.$

indicating that the total reward, f_n, at stage n is equal to the reward available at stage n-1, obtained as the result of the decision made previously and the state of the process at stage n. Recursion relations can be considered as a mathematical expression of an intuitive principle known as the "principle of optimality" first isolated and stated by Bellman (1957, p. 83) but reported more crudely here as a restatement of the definition of recursion relations.

3.1.1 Backward and Forward Recursion

Dynamic Programming Analysis is based on the mathematical concept of recursion. It essentially transforms a problem in which an optimal policy of n decisions is desired, to a sequential decision process of n separate stages dealt with in a recursive manner. The recursion may take place either backward or forward within the process. Sometimes it is more convenient (or essential) to proceed one way or the other. The basic difference is that in forward recursion the analysis proceeds from the last stage to the first. In backward recursion the analysis proceeds from the first stage to the last.

For many multi-stage decision problems there is no need to distinguish backward and forward solutions and the choice employed is often a matter of personal preference.

3.2 Optimality Principle

The optimal reward obtainable at stage n is equal to the reward available at stage n, derived as the result of making that decision, conditional on the process state at stage n, which optimizes the sum of the immediate reward and the reward from previous stages. Bellman states it thus:

"An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must consitute an optimal policy with regard to the state resulting from the first decision."

I have always believed that this statement was more meaningful (to me) if the words "initial" and "first" were replaced by the word "previous." This seems more appropriate for MSDP rather than two stage processes. Aris (1965) describes the principle as:

"If you don't do the best you can with what you happen to have got, you will never do the best you might have done with what you should have had."

Stated even more simply the principle tells us that if we know how to proceed with the best course of action from any intermediate state to the final state and stage, we should follow this course of action regardless of how we arrive at the intermediate state and stage from an initial state and stage.

Dynamic Programming provides a systematic procedure for determining what information is required and how it should be effectively used for optimizing the decisions required by a multi-stage decision process.

3.3 Transition

Each stage of the process transforms its input state into an output state (for that stage) in a way dependent on the decision made for the operation of that stage.

3.4 Formulation of Dynamic Programming Analysis

Many who use DP to optimize MSDP find it useful to formulate the information about the process in a relatively structural way. We now indicate this structure somewhat generally and intend to detail it more specifically in a later section. The structure has five phases which we describe as:

I. Problem Statement Phase. Here the problem and solution goal are described using natural and operational language.

II. The multi-stage decision process which is related to the information given in Phase I is characterized as a multi-stage decision process by defining stage, state, decision, constraints and rewards in terms of the description given in Phase I.

III. The dynamic programming analysis is characterized for the MSDP of the problem statement by specifying the recursion relations, the state transformation, the state space, the decision space, the state variable and the decision variable as well as the optimization type.

IV. The approach to be used to compute the optimal policy is defined and the calculations are carried out.

V. The optimal policy is presented on operational terms and any pertinent information relative to the analysis or the problem solution are provided.

3.5 Methods of Solution

Most methods of finding an optimum policy by Dynamic Programming Analysis begin with the recursion relation equations and treat these in various ways. These techniques we describe as either analytical or tabular.

3.5.1 Analytic Solutions

This method of solution proceeds mathematically with the goal of obtaining separate and multiple functions of the problem's variables that can be obtimized simply and independently by mathematical means such as differential calculus, rather than optimizing the recursion relations directly as a single function of multiple variables. Tiechroew (1964, p. 610) gives a simple example of this process for an inventory problem.

Solutions may also be obtained by complete enumeration, heuristic procedures, simulations and by using Network Routing analysis, see Exhibit 1.

3.5.2 Tabular Solutions

This method is especially adapted to computer solutions and proceeds by enumerating stage-wise optimum solutions and organizing these in a table or tables as the process proceeds. The tables are portrayed and from them the optimal policy is determined.

We shall employ, later, a particular tabular method that we call the "Diagonal Method" to solve the allocation problem used for illustration in this presentation. Each class of problem to which DP is applied usually requires a special type of analytical or tabular solution so that no general solution algorithm exists for DP as it does for Linear Programming with the Simplex. We shall find the Diagonal Method of solution easy to carry out.

3.6 The Dynamic Programmers

3.6.1 The Theorists

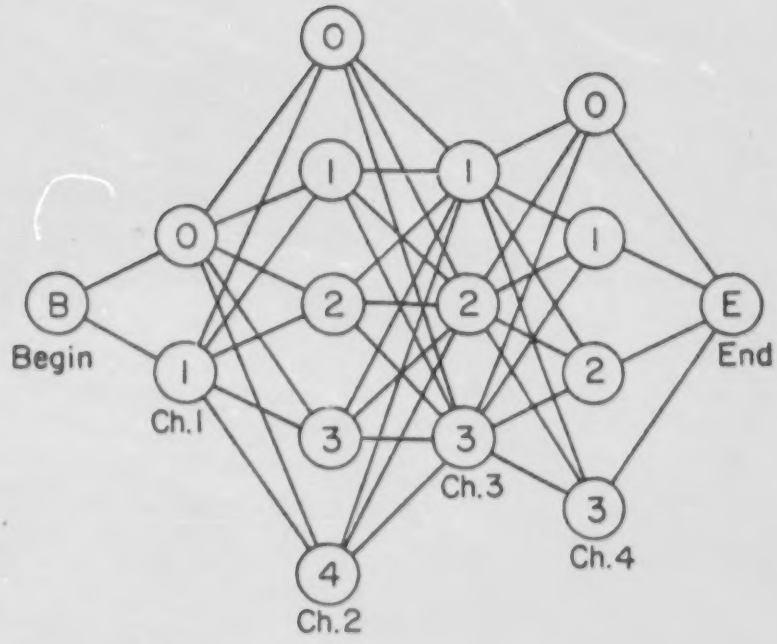
Readers of this report have no doubt detected the author's boundless admiration for Professor Richard E. Bellman and his mathematical-scientific work. Let me relate a few pertinent facts.

1. He was born in N.Y.C. in 1926.

2. He received his BA in 1941 from Brooklyn College.

EXHIBIT 1

The Curriculum Problem Portrayed as a Network Routing Problem



He received his MA in 1943 from the University of Wisconsin.

He received his Ph.D in 1946 from Princeton University.

3. He is, by education and profession, a Mathematician.

4. He has worked at the following locations:

   '44-'45 Radio Sound Laboratory (NEL) in San Diego, CA

   '45-'46 Los Alamos, NM

   '46-'47 Princeton, NJ

   '47-'52 Stanford University

   '52-'70 RAND Corp.

   '70-Present University of Southern California

5. In his present capacity at USC he holds professorships simultaneously in the fields of medicine, electrical engineering and mathematics!

6. As of about 1968 he had written 15 books and 400 professional reports and articles.

7. He has originated and developed the mathematical fields of Dynamic Programming and Automatic Control Theory. A photograph of him may be found in Exhibit 15.

To appreciate this obvious genius of our time, we need only compare our own personal biography with the seven points mentioned above. It seems apparent that Dr. Bellman's mathematical work represents outgrowth of mathematical edifices constructed by Wald (1950) and von Neumann (1947) and that he represents and extends their traditions. We should also mention S. Dreyfus, (1962) a numerical analyst, and Dr. R. Kalaba, a mathematician, both of whom worked and published with Bellman while he was at RAND. Meanwhile, at Harvard, another theoretician, Professor Ronald Howard (1960) was contributing to and publishing in the field of DP.

3.6.2 The Literature

There are approximately 2000 references on the theory and application of Dynamic Programming. Up to 1968 there were only a handful of good textbooks on the subject which we could identify from among the following:

Aris (1964, 1965)
Bellman (1957)
Bellman and Dreyfus (1962)
Bellman and Kalaba (1965)
Howard (1960)
Hadley (1964)
Kaufmann (1967)
Nemhauser (1966)
Roberts (1964)

Other literature was dispersed in RAND publications, technical journals such as JORSA and Management Science. This made things difficult for students, particularly those for whom mathematical logic did not come easily. More recently, however, the available books or texts or chapters on DP has increased from a handful to an armful. We mention a few more:

Cooper (1974)
Dano (1975)
Hastings (1973)
House (1972)
Shamblin and Stevens (1974)
Wagner (1969)
White (1969)

So the student of DP now has a choice for the style of learning he wishes to pursue. Bellman's original books remain basic to the subject nevertheless. Moreover, many texts on Operations Research include chapters on DP.

3.7 Some Applications

The literature cited herein references numerous applications of Dynamic Programming to multi-stage decision processes occurring in applied situations. Bellman and Dreyfus (1962) give many computational illustrations.

Allocation problems are known by a wide variety of special names such as the Knapsack Problem, the Assignment Problem, etc. These are amply illustrated in the literature. As a 10th grade student my son, Moonan (1968) applied Dynamic Programming to the problem of determining the optimal policy for teaching reading using a Markov process and a DP procedure suggested by Howard (1960).

Wagner (1969) gives examples of the Distribution of Effort Problem whereas Page (1975, 1976) provides the only known example of the curriculum problem analyzed by Dynamic Programming.

The author, Moonan and Covher (1966), applied Dynamic Programming to the problem of determining an optimal promotion policy for men in a Navy rating.

4. ALLOCATION PROBLEMS

4.1 The General Allocation Problem

It seems to this author that personnel management and personnel research activities abound with a class of problems known as "Allocation Problems." Such a problem occurs when a limited supply of one or more resources such as raw material, money, time, manpower, space, capacity, etc. is required to be allocated among a number of uses in such a way as to optimize some return related to the allocations. Often the allocation activities (the personnel assignment problem, for example) represent multi-stage decision processes or can be so represented. In Exhibit 2 we characterize a general allocation problem in MSDP terms and in Exhibit 3 express it in DP terms.

An allocation problem occurs when there exists some kinds or amounts of resources which are required as inputs to activities which produce some kind of outputs. These resources often have alternative uses, and costs are sustained in using a resource. Similarly, there is a resultant benefit obtained from using a resource.

Solution of the allocation problem occurs when the allocations are determined which either minimize the costs associated with a given benefit, or which maximize the benefit for a given cost.

4.2 The Distribution of Effort Problem

When the allocations are forms of allotments, dispersions, partition or similar apportionment, and the resource is related to labor, work, expenditure of effort or similar performance, the Allocation Problem is sometimes referred to, Wagner (1969) as the "Distribution of Effort Problem."

4.3 The Curriculum Problem

The Curriculum Problem is an allocation problem or Distribution of Effort Problem occurring in the field of education and brilliantly identified by Page (1975) as a multi-stage decision problem, properly optimized by the Methods of Dynamic Programming.

Briefly stated the problem goes something like this:

An instructor of a certain school subject has a variety of learning activities available for his students. Each activity may utilize a certain amount of the student's time which is in limited supply and depending upon his intensity of devotion to the activity, this effort yields, in the teacher's mind, some educational benefit.

Page's (1976) example used textbook chapters as the activities to be participated in although this is just one educational illustration. Before Page's work, solutions to these problems were, for the most part, informally obtained and were often unsatisfactory.

EXHIBIT 2

The Allocation Problem Expressed in Multi-Stage

Term	Definition	Symbol
Stage	The allocation of a resource to a use n=nr· of uses remaining to receive allocation	n
State	The amount of resource remaining to be allocated and n uses remain to be considered	(n,x)
Decision	Allocate d amount to the n-th stage use	d
Reward	The reward obtained from allocating resource in amount d_n to use n	$r(n,d_n)$

EXHIBIT 3

The Allocation Problem Expressed in Dynamic Programming Terms

Term	Definition	Symbol
State Variable	The amount of resources remaining to be allocated to a use	x
Decision Variable	Amount of resource allocated to use	d
Transition	Allocation of amount d at current stage leaves amount x-d available at next stage	$(n,x)\rightarrow(n-1,x-d)$ or $(n,x)\rightarrow(n+1,x-d)$
Stage Space	Let N be the number of uses to possibly receive an allocation of the resource. The stage space then is	$o\leq n\leq N$
State Space	Let X be the number of units of resource available for allocation. The state space then is	$o\leq x\leq \bar{X}$
Decision Space	Let D_n be the maximum number of units of resource available for allocation to use n. The Decision Space then is	$o\leq d\leq D_n, n=1,\ldots,N$
Optimal State Value	The total reward when state (n,x) is the starting state and optimal decisions are made with n uses remaining to be considered	$f(n,x)$
Recursion Relation	d is the set of feasible allocations associated with state (n,x)	$f(n,x)=$ $op_d t[r(n,d)+$

4.4 Methods of Reward Determination

A difficult problem which occurs to many decision makers in applying Decision Theory in operational situations is the problem of specifying the "rewards," "payoffs," "return," etc. expected to accrue as the result of a decision. This problem is especially acute in those situations where the reward cannot easily be expressed in monetary terms. Such is the case for many allocation problems and in particular for the curriculum problem.

4.4.1 Token Method

Page (1976) has recognized the necessity of incorporating the value system of the decision maker into the decision making process and recommends determining the educational value of a student's level of effort, for each activity, by the use of the "token strategy" see Page (1972, 1974, 1976). We will use the resulting numbers as rewards in our example.

4.4.2 The Probability Ratio Method

The author has, in the course of using Decision Theory, developed his own valuing procedures - the latest being called PRM, the Probability Ratio Method. This method has the advantage of being quickly and easily applied and is detailed in Moonan (1976).

The procedure requires the decision maker to supply his subjectively determined probability that each item of the set of items to be valued, if realized, will satisfy the goal for which the decision is being made. For the problem at hand, the "items" are the educational result obtained by the student from studying a chapter with a given level of effort.

These probabilities are then equated to a "utility model" which incorporates a personal factor parameter of the decision maker and a value factor of the item. Using simple probability theory the values can then be estimated for each item independently of the value of the personal factor parameter of the decision maker.

It seems to this author that the personal factor is required by all methods used to estimate the subjective value of something relative to a particular person, because if this factor measures anxiety for goal achievement. This state influences his judgment regarding the value of "solutions." The PRM recognizes this fact and then proceeds, by mathematics, to estimate the values independently of the personal factor.

4.4.3 Other Methods

Solutions to the value determination problem which, in principle, could be used for determining rewards for the curriculum problem have been suggested by von Neumann and Morgenstern (1947), their

method being called the "standard gamble technique," Raiffa (1968), and Churchman, Ackoff, and Arnoff (1957), and Fishburn (1964).

5. SOLVING THE CURRICULUM PROBLEM

In section 4.3 we discussed the nature of the curriculum problem of education. We next well illustrated how the optimum policy for this problem is achieved using Dynamic Programming. In order to do this we shall use the example provided by Page (1975, 1976).

5.1 The Problem

A teacher of a college subject desires that the students of his class use their available time of X=10 study hours in a week devoted to each of 4 chapters of the textbook. Each chapter may be studied differentially by the student's applying one of five levels of intensity of effort to each chapter. These decisions are defined as follows:

		Acronyms
Decision 1:	No study given the chapter	Omit
Decision 2:	Light, once over reading of the chapter	Read Once
Decision 3:	Intensive reading for understanding	Read Intensively
Decision 4:	Some problem solution	Do Some Problems
Decision 5:	Solving all problems	Do All Problems

Utilizing a particular level of effort (decision) on a chapter expends part of the student's 10 hours of available study time, but this expenditure has its own educational reward.

The problem to be solved is: "What strategy of study should be employed by the student in order to achieve the maximal educational reward for 10 hours of study?" Stating this otherwise we might ask, "What is the student's optimal study policy?" Study policy being the set of decisions selected for studying each chapter.

5.2 The Data

The information used for solving this problem is shown in Exhibit 4. This information shows the amount of study time expenditure and the associated educational reward r expected for each level of study (decision) of each chapter. Note that only for the second chapter are all five levels of effort admissible. Apparently, study of the first chapter for more than three hours has no additional benefit.

EXHIBIT 4

Study Time, x_1 and Educational Reward, r_1
for 4 Chapters and 5 Levels of Effort

Chapter	x=time r=reward	Level of Effort 1	2	3	4	5
1	x_1	0	3	-	-	-
	r_1	0	9	-	-	-
2	x_2	0	2	4	7	8
	r_2	0	8	12	15	17
3	x_3	-	3	4	6	-
	r_3	-	7	8	9	-
4	x_4	0	2	5	7	-
	r_4	0	10	12	15	-

Note: The - indicates the level of effort is inappropriate. The heavy lines delineate the admissible levels of effort for all chapters.

The educational rewards were determined by Page's token technique. While at NPRDC he developed a system for obtaining these values during group consultations.

5.3 Multi-Stage Decision Process Aspects of the Curriculum Problem

We define, for convenience, the Curriculum Problem in terms of a multi-stage decision process as shown in Exhibit 5. These definitions are determined in a straight forward manner from the problem statement.

5.4 Dynamic Programming Analysis Aspects of the Curriculum Problem

Our mode of proceeding with the determination of optimum policies for allocation problems has been first to consider their multi-stage decision process properties, and then to express the features required by dynamic programming analysis to determine their optimum policy. In Exhibit 6 we perform the latter task for the Curriculum Problem.

5.5 A Solution Procedure

Our problem is to distribute 10 hours of study time among four chapters, subject to conditions shown in Exhibit 4 in such a manner that the educational reward to the student will be maximized. Stated in this way we can see Page's description of the problem is idealized, since the true educational problem probably requires Exhibit 4 to be individualized for each student and each would perhaps have a different total study time available. Nevertheless, let us learn to walk before we run, and solve the given problem.

Of the many solution procedures available to us let us arbitrarily select a tabular method.

We do this because the four recursion relations

$$(2)\quad f(1,x)=\underset{d(k)}{\text{opt}}\,[r(1,k)+f(o,x)] \text{ where } f(o,x)=o$$

$$(3)\quad f(2,x)=\underset{d(k)}{\text{opt}}[r(2,k)+f(1,x)]$$

(X)

$$(4)\quad f(3,x)=\underset{d(k)}{\text{opt}}[r(3,k)+f(2,x)]$$

$$(5)\quad f(4,x)=\underset{d(k)}{\text{opt}}[r(4,k)+f(3,x)]$$

are so straight forward and simple that some simple tabular procedure will surely be satisfactory and, perhaps, instructive. Moreover, we may wish to computerize the solution as Moonan (1976) and want to keep matters easy for the computer programmers. We partition our task accordingly into three programming phases called

EXHIBIT 5

The Curriculum Problem Expressed in Multi-Stage Decision Process Terms

Term	Definition	Symbol
State	A chapter to be studied	n
State	The amount of unexpended study time	x
Decision	The level of effort k used to study chapter n	d(n,k)
Reward	The educational benefit derived from studying chapter n with level of effort k	r(n,k)

EXHIBIT 6

The Curriculum Problem Expressed in Dynamic Programming Analysis Terms

Term	Definition	Symbol
State Variable	A Chapter to be studied by a student	
State Variable	x Amount of study time remains to be utilized and n chapters remain to be studied	
Decision Variable	Apply level of effort d_k to current chapter.	$d_k d(k)$
Allocation	Amount of time used for level of effort k.	$t(n,k)$
Transition	Study of the current chapter reduces available time for next chapter.	$(n,x) \rightarrow (n-1, x-t(n,k))$
Stage Space	Number of chapters to be studied	$0 \leq n \leq 4$ chapters
State Space	The study times available	$0 \leq x \leq 10$ hours
Decision Space	Levels of effort available for study	$d(n,k)$, and $1 \leq k \leq 5$
Optimal State Value	Total Educational reward when starting state is (n,x) and an optimal plan is followed.	
Recursion Relation	Backward Recursion	$f(n,x) = \underset{d}{\text{opt}}[r(n,k)+f(n-1,x)], n = 1,2,3,4,$

a) Input,

b) Process, and

c) Output.

5.5.1 The Input Phase

i) Title Card: Gives Title and Date of Analysis

For Distribution of Effort (DOE) problems we have only two other types of input. These are

ii) Parameters

K=nr of possible levels of effort (decisions)

N=nr of activities

T=type of optimization

T=1, maximization of rewards

T=0 minimization of costs

X=Total resource available for allocation

iii) Data

The data of DOE problems are the amounts of resource to be allocated by activity and decision, and the associated rewards or costs. For our problem:

K=5
N=4
T=1
X=10, and the data are given in Exhibit 4.

Additionally, we input K cards containing the decision number and its acronym.

5.5.2 The Process Phase

Our programming plan involves the creation of a new table which presents the data of Exhibit 4 in more extended form. This table is shown in Exhibit 7. This work is not really necessary since we could use Exhibit 4 in the subsequent calculations, but we do construct it only for the sake of convenience to the programming process.

Our plan also involves evaluating (3) in tabular form and then successively (4) and (5) using other tables. It is not strictly necessary to formally construct these tables, but we do so for the sake of calculational clarity.

EXHIBIT 7

The Extended Form of the Input Data

Chapter	1		2		3		4	
x	r(1,k)	d(k)	r(2,k)	d(k)	r(3,k)	d(k)	r(4,k)	d(k)
0	0	1	0	1	-	-	0	1
1	0	1	0	1	-	-	0	1
2	0	1	8	2	-	-	10	2
3	9	2	8	2	7	2	10	2
4	-	-	12	3	8	3	10	2
5	-	-	12	3	8	3	12	3
6	-	-	12	3	9	4	12	3
7	-	-	15	4	-	-	15	4
8	-	-	17	5	-	-	-	-
9	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-

Our immediate purpose then is to evaluate f(2,x) of (3). To do this we see from (2) that f(1,x)=r(1,k) because f(o,x)=o. This allows us to write (2) as

(6) $$f(2,x)=\underset{d(k)}{\text{opt}}[r(2,k)+r(1,k)]$$

the functional values of which are conveniently expressed in tabular form to be shown shortly. We build this table somewhat like a coordinate system whose abscissa vectors are X, r(1,k) and d(k), the X vector being X=0,1,2,...,X=10 and r(1,k) and d(k) as obtained from Exhibit 7. Notice we have placed these three vectors at the top of our coordinate table, rather than at the bottom which is usual, see Exhibit 8. Similarly, the ordinate of the table is associated with the three vectors X, r(2,k) and d(k). Having established the bordering vectors we now compute r(2,k)+ r(1,k) as required by (6) for the admissible X of Chapters 1 and 2. For Chapter 1, 0<X<3 are admissible study times and for Chapter 2, 0<X< 8 are admissible. The totals of r(1,k) and r(2,k) are shown in Exhibit 9.

We now must declare the maxima of these sums, for given X, as required by (3). Suppose X=3, study times of X_1=0,1,2,3 in combination with study times of X_2=0,1,2,3 can be arranged in pairs so that the total is X=3. The rewards for these combinations appear in Exhibit 9 on the diagonal alls (X_1,X_2)=(0,3),(1,2),(2,1) and (3,0) giving possible values of f(2,3)=9,088. The maximum of which is 9 so we set f(2,3)=9. The diagonals associated with other X in Exhibit 9 are shown marked in Exhibit 10 and the values of f(2,X) are given in Exhibit 11 and we have achieved our intermediate computational goal. Diagonal maxima are indicated with *'s.

Our next problem is to evaluate (4) and (5). For (4) we construct a table similar to Exhibit 9 using X,r(3,k),d(3) as abscissa vectors, and X,f(2,X)d(2) as ordinate vectors. The result is shown in Exhibit 12 and f(3,X) are given in Exhibit 11.

Similarly for (5) we build Exhibit 13 with X,r(4,k),d(4) as abscissa vectors and with X,f(3,X),d(3) as ordinate vectors, with f(4,X) given in Exhibit 11.

This process could be repeated for any number of Chapters (Activities) in the problem. We have but one remaining step in our processing phase of the program. This step requires us to prepare Exhibit 14 for output. The information for this table comes from Input and Exhibit 11. Its construction completes the Process phase of our program.

5.5.3 The Output Phase

We suggest that the following information be given as output, in this order:

EXHIBIT 8

Partial Table for Computing f(2,x)

			d_x	1	1	2	2	3	3	3	4	5	-	-
			R_x	0	0	8	8	12	12	12	15	17	-	-
	d_x	R_x	X	0	1	2	3	4	5	6	7	8	9	10
	1	0	0											
	1	0	1											
	1	0	2											
	2	9	3											
	-	-	4											
	-	-	5											
	-	-	6											
	-	-	7											
	-	-	8											
	-	-	9											
	-	-	10											

EXHIBIT 9

Table for Computing f(2,x)

			d_x	1	1	2	2	3	3	3	4	5	-	-
			R_x	0	0	8	8	12	12	12	15	17	-	-
	d_x	R_x	X	0	1	2	3	4	5	6	7	8	9	10
	1	0	0	0	0	8	8	12	12	12	15	17	-	-
	1	0	1	0	0	8	8	12	12	12	15	17	-	-
	1	0	2	0	0	8	8	12	12	12	15	17	-	-
	2	9	3	9	9	17	17	21	21	21	24	26	-	-
	-	-	4	-	-	-	-	-	-	-	-	-	-	-
	-	-	5	-	-	-	-	-	-	-	-	-	-	-
	-	-	6	-	-	-	-	-	-	-	-	-	-	-
	-	-	7	-	-	-	-	-	-	-	-	-	-	-
	-	-	8	-	-	-	-	-	-	-	-	-	-	-
	-	-	9	-	-	-	-	-	-	-	-	-	-	-
	-	-	10	-	-	-	-	-	-	-	-	-	-	-

EXHIBIT 10

Table for Computing f(2,x) With Diagonals

			d_x	1	1	2	2	3	3	3	4	5	-	-
			R_x	0	0	8	8	12	12	12	15	17	-	-
	d_x	R_x	X	0	1	2	3	4	5	6	7	8	9	10
	1	0	0	0*	0*	8*	8	12*	12	12	15	17	-	-
	1	0	1	0*	0	8	8	12	12	12	15	17	-	-
	1	0	2	0	0	8	8	12	12	12	15	17	-	-
	2	9	3	9*	9	17*	17*	21*	21*	21*	24*	26	-	-
	-	-	4	-	-	-	-	-	-	-	-	-	-	-
	-	-	5	-	-	-	-	-	-	-	-	-	-	-
	-	-	6	-	-	-	-	-	-	-	-	-	-	-
	-	-	7	-	-	-	-	-	-	-	-	-	-	-
	-	-	8	-	-	-	-	-	-	-	-	-	-	-
	-	-	9	-	-	-	-	-	-	-	-	-	-	-
	-	-	10	-	-	-	-	-	-	-	-	-	-	-

*Diagonal Maxima

EXHIBIT 11

Optimal State Values for Stages 2, 3, and 4

x	f(2,x)	f(3,x)	f(4,x)
0	0	-	-
1	0	-	-
2	8	-	-
3	9	7	7
4	12	8	8
5	17	15	17
6	17	16	18
7	21	19	25
8	21	24	26
9	21	25	29
10	24	28	34

EXHIBIT 12

Table for Computing f(3,x)

			d_x	-	-	-	2	3	3	4	-	-	-	-
			R_x	-	-	-	7	8	8	9	-	-	-	-
	d_x	R_x	X	0	1	2	3	4	5	6	7	8	9	10
		0	0	-	-	-	7*	8*	8	9	-	-	-	-
		0	1	-	-	-	7	8	8	9	-	-	-	-
		8	2	-	-	-	15*	16*	16	17	-	-		
		9	3	-	-	-	16*	17	17	18	-			
		12	4	-	-	-	19*	20	20	21				
		17	5	-	-	-	24*	25*	25					
		17	6	-	-	-	24	25						
		21	7	-	-	-	28*							
		21	8	-	-	-								
		21	9	-	-	-								
		24	10	-	-	-								

*Diagonal Maxima

EXHIBIT 13

Table for Computing f(4,x)

			d_x	1	1	2	2	2	3	3	4	-	-	-
			R_x	0	0	10	10	10	12	12	15	-	-	-
	d_x	R_x	X	0	1	2	3	4	5	6	7	8	9	10
		-	0	-	-	-	-	-	-	-	-	-	-	-
		-	1	-	-	-	-	-	-	-	-	-	-	-
		-	2	-	-	-	-	-	-	-	-	-	-	-
		7	3	7*	7	17*	17	17	19	19	22			
		8	4	8*	8	18*	18	18	20	20				
		15	5	15	15	25*	25	25	27					
		16	6	16	16	26	26	26						
		19	7	19	19	29*	29							
		24	8	24*	24	34*								
		25	9	25	25									
		28	10	28										

*Diagonal Maxima

EXHIBIT 14

MSDP Optimal Allocations

Activity	Decision	Action	Allocation	Reward
1	2	Read Once	3	9
2	2	Read Once	2	8
3	2	Read Once	3	7
4	2	Read Once	2	10
Totals			10	34

1. The Title Card

2. The Parameters: K, N, T, $\bar{X}$

3. Exhibit 4

4. Exhibit 11

5. Exhibit 14

6. SUMMARY

We have attempted a difficult, but important, task, namely, to discuss rather descriptively a complex mathematical topic - namely, Dynamic Programming Analysis and its application to a class of multi-stage decision processes commonly referred to as allocation problems. Our procedure has involved, as much as possible, the separation of the multi-stage decision aspects of an allocation problem from the dynamic programming aspects of optimization of the decisions required. This was done in the hope that clarification of analysis would be increased and complexity would be reduced thereby.

It was asserted, with example, that allocation problems occur frequently in the course of military personnel research and management. A particular allocation problem known as the Distribution of Effort was identified as the Curriculum Problem of Education and this problem was solved by Dynamic Programming Analysis using the structures, notation and techniques outlined in the report.

EXHIBIT 15

The Decision Theory Mathematicians

A PERFECTER OF QUESTIONS
Poring over figures in the computer laboratory *(right)* is the RAND Corporation's athletic Dr. Richard Bellman, who plays one of the most furious games of tennis in California. His mathematical specialty is dynamic programming, which sets up a computer so that it can keep finding answers as the problem it is working on changes. Typical examples are the minute-to-minute guidance of a satellite or day-to-day analysis of the fluctuating stock market. The difficulty, says Dr. Bellman, is in formulating a problem. "We often find," he says, "that a good question is more important than a good answer."

A GIANT OF OUR TIME
John Von Neumann, whose mathematical "game theory" is now used in business decision making and cold-war strategy, was famed for solving in his head problems that made other mathematicians turn to pencil and paper or even a calculator. Colleagues sometimes wondered, half-seriously, if his "lightning quick" intellect did not suggest "a species superior to man." His unique mind did, at least, set him apart from most men: after an automobile crash at Princeton, New Jersey, for example, he emerged from his wrecked car to explain, "The trees on the right were passing me in orderly fashion at 60 miles per hour. Suddenly one of them stepped out in my path!"

ABRAHAM WALD

who contributed vitally to the advancement of mathematical statistics through his broad and fundamental research which will continue to influence the development of statistical theory and practice and who will long be remembered as an inspiring and esteemed teacher and colleague.

474

7. REFERENCES

Aris, R. Discrete dynamic programming. New York: Blaisde[illegible], 1964.

Aris, R. Dynamic programming. Socony Mobil Oil Company, [illegible]65.

Bellman, R. Dynamic programming. Princeton: Princeton [illegible]niversity Press, 1957.

Bellman, R. Adaptive control processes: A guided tour. Princeton: Princeton University Press, 1961.

Bellman, R., & Dreyfus, S. Applied dynamic programming. Princeton: Princeton University Press, 1962.

Bellman, R., & Kalaba, R. Dynamic programming and modern control theory. New York: Academic Press, 1965.

Bergamini, D. Mathematics (p. 59). New York: Life Sciences Library, Time Inc., 1963.

Churchman, A. A. Introduction to operations research. New York: John Wiley & Sons, 1957.

Cooper, L. Applied nonlinear programming for engineers and scientists. Englewood, NJ: Aloray, Inc., 1974.

Dano, S. Nonlinear and dynamic programming. New York: Springer-Verlag, 1975.

Fishburn, P. Decision and value theory. New York: John Wiley & Sons, 1964.

Hadley, G. Nonlinear and dynamic programming. Reading, MA: Addison-Wesley, 1964.

Hastings, N. Dynamic programming (with management applications). New York: Russak & Company, 1973.

House, W. Operations research. Princeton: Auerbach, 1972.

Howard, R. Dynamic programming and Markov processes. New York: John Wiley & Sons, 1960.

Kaufmann, A. Graphs, dynamic programming, and finite games. New York: Academic Press, 1967.

Moonan, W. K. Using a Markov process to determine reading instruction policies. Proceedings of the Greater San Diego Science Fair, 1968.

Moonan, W. J. FUTILE: Fast measurement of consequence utility using a logistic deliberation model with decision theory: Tools of analysis and decision (FACT DOCUMENT No. 51). San Diego: Navy Personnel Research and Development Center, 1976.

Moonan, W. J., & Covher, M. A computer program for the determination of an optimal advancement policy for petty officers: A dynamic programming approach. San Diego: Navy Personnel Research and Development Center, 1966.

Moonan, W. J., & Covher, M. DOE: Distribution of effort: Tools for analysis and decision (FACT DOCUMENT No. 51. San Diego: Navy Personnel Research and Development Center, 1976.

Nemhauser, G. Introduction to dynamic programming. New York: John Wiley & Sons, 1966.

Page, E. B. Seeking a measure of general educational advancement: The bentee. Journal of Educational Measurement, 1972, 9, 33-43.

Page, E. B. "Top-down" trees of educational values. Educational and Psychological Measurement, 1974, 34, 573-584.

Page, E. B. Design of Navy course structure through a dynamic programming algorithm. San Diego: Navy Personnel Research and Development Center, 1975.

Page, E. B. Curriculum design through operations research. American Educational Research Journal, 1976, 13(1), 31-49.

Raiffa, H. Decision analysis introductory lectures on choices under uncertainty. Reading, MA: Addison-Wesley, 1968.

Roberts, S. Dynamic programming in chemical engineering and process control. New York: Academic Press, 1964.

Shamblin, J. E., & Stevens, G. T., Jr. Operations research, a fundamental approach. New York: McGraw-Hill, 1974.

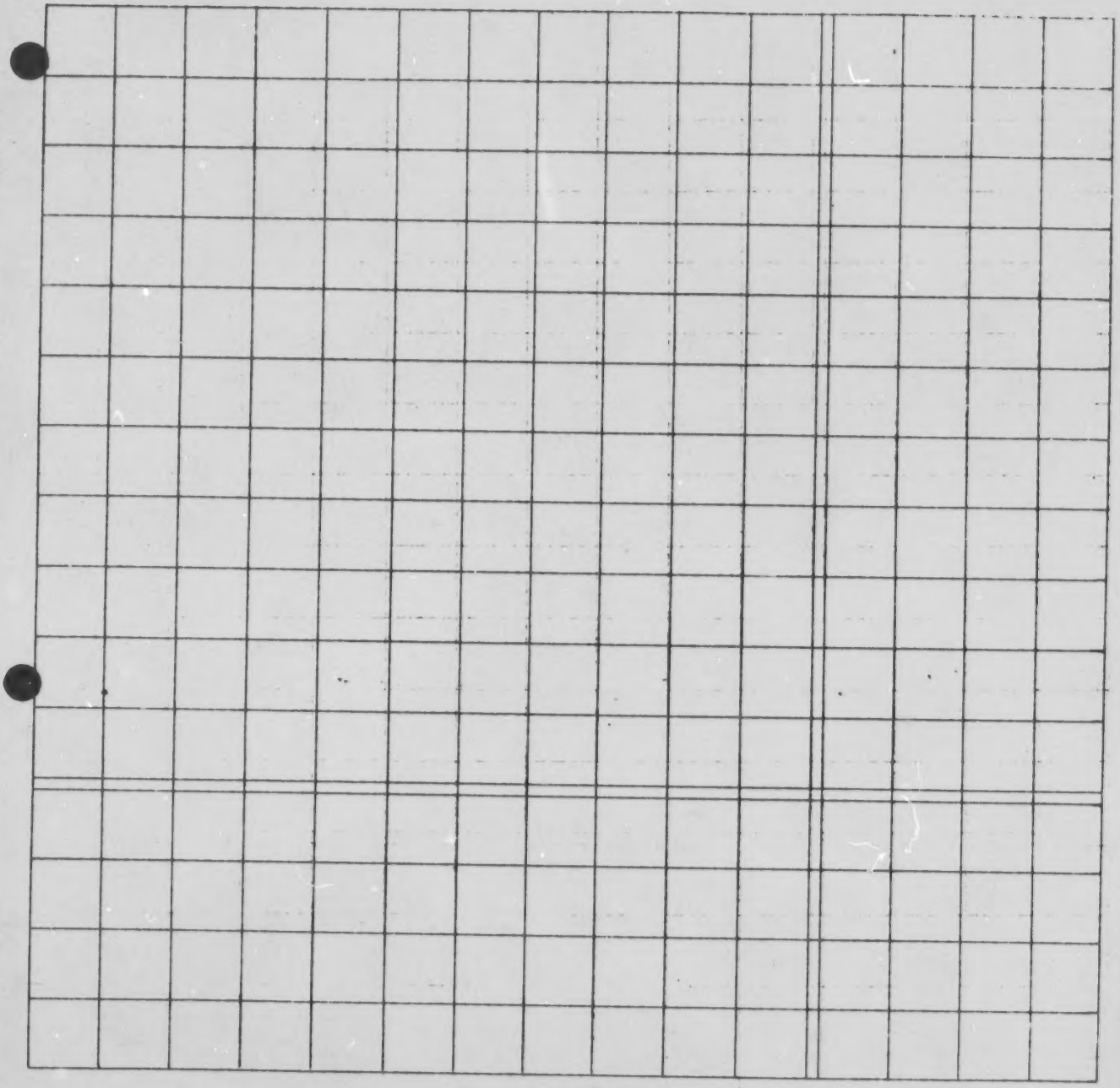
Teichroew, D. An introduction to management science deterministic models. New York: John Wiley & Sons, 1964.

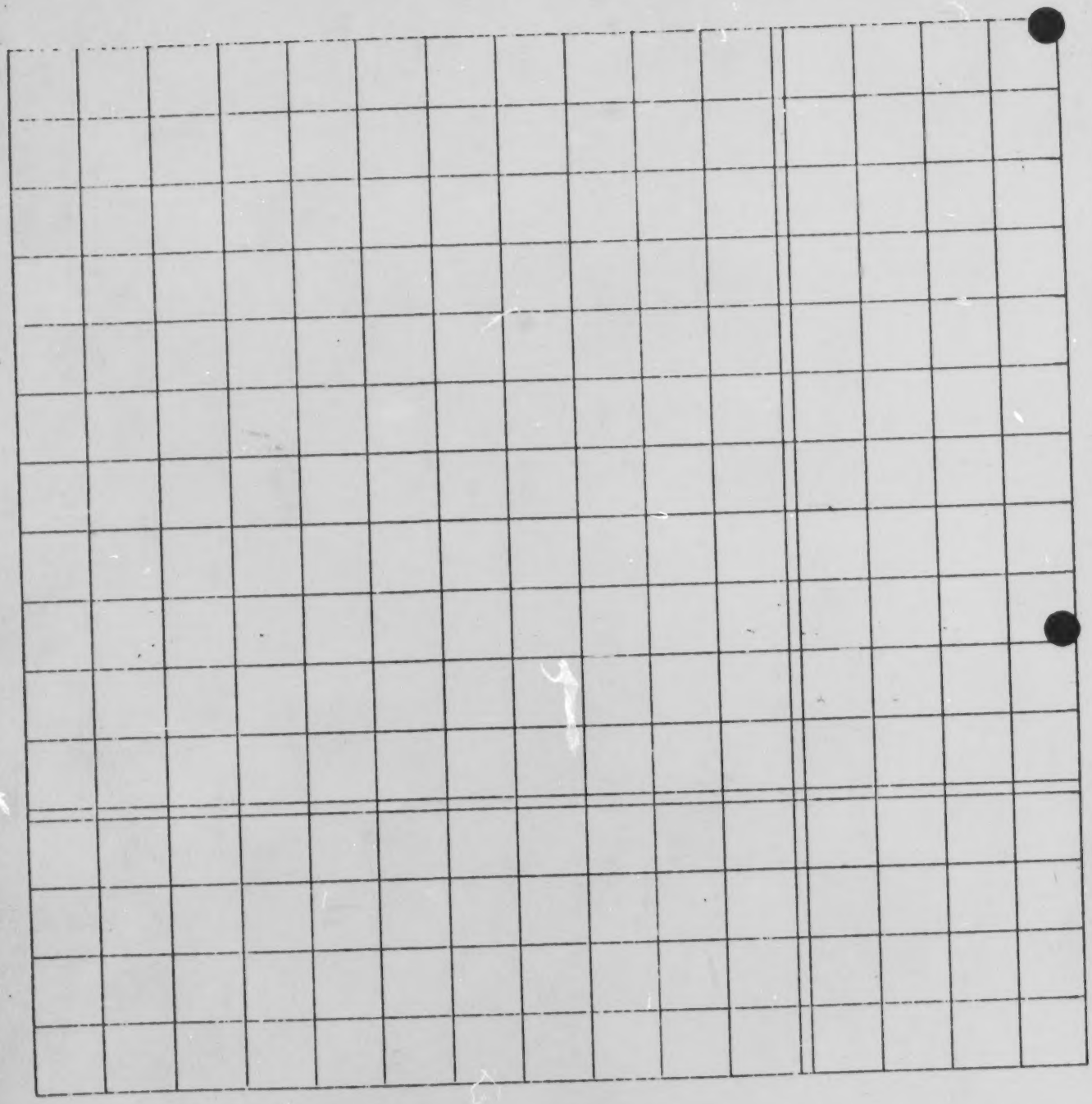
Von Neumann, J., & Morgenstern, O. Theory of games and economic behavior. Princeton: Princeton University Press, 1947.

Wagner, H. M. Principles of operations research: With applications to managerial decisions. Englewood Cliffs, NJ: Prentice-Hall, 1969.

Wald, A. Statistical decision functions. New York: Chelsea Publishing Company, 1950.

White, D. J. Dynamic programming. San Francisco: Holden-Day, 1969.





EFFICACIES OF THREE METHODS OF WEIGHTING BIODATA TO PREDICT TURNOVER

Warren S. Blumenfeld

Georgia State University

The armed forces recently went to a total volunteer concept. The draft, as it has been known for the last 30+ years, was halted. For purposes of inducing individuals to enter the military, several new options are being offered. Recruiting costs per individual have gone up substantially as a function of offering new incentives and fringe benefits, as well as increasing the numbers of personnel whose function is recruiting. The overall investment on the part of the government (management) for each new selection risk has likewise gone up considerably. For every individual who does not complete contract obligation, management suffers directly and indirectly, not only in economic cost, but in terms of planning and manpower utilization. Therefore, it would obviously be to the advantage of management in the areas of planning, staffing, and training to improve their "batting average" in terms of the recruiting and selection of individuals who will complete military contract obligation.

As direct and indirect personnel costs continue to rise in order to "sell" the military (or "buy" personnel), recruiting will no doubt become an increasingly more important personnel function. When the draft was available to induce individuals to enlist, recruiting was of less importance. Now, marketing (and "prospecting" in particular) will be of greater importance. However, it is possible that recruiting costs can be better managed if most of the effort and money is spent to attract those individuals who are most likely to complete their military contract obligation.

The task becomes one of identifying the characteristics of those individuals for whom the likelihood of completing their contract obligation is greatest. For recruiting purposes, the "key" to a weighted application blank could give the organization a (verbal) behavioral profile of the kind of individual they might best seek. Such a market segmentation behavioral profile, if successfully developed, could aid management in their recruiting and selection decisions.

Weighted Application Blanks

A body of literature has developed in applied psychology regarding the development of weighted application blanks for the prediction of various criteria, mostly of a personnel type. See, for example, Blumenfeld (1973). Essentially, this involves: (a) the development of a scoring key to translate qualitative, usually non-cognitive information, i.e., biodata, into quantitative terms, and (b) the subsequent application of the individual differences described in quantitative terms toward the prediction of the dependent variable of interest. Appropriate and inappropriate experimental designs have been indicated; see, for example, Blumenfeld (1972), Blumenfeld, Godbey, and Blumenfeld (1973), Cureton (1950), Dunnette and Kirchner (1962), and Kurtz (1948).

Some illustrative examples in which the application of this paradigm has been accomplished successfully in the area of personnel research include occupational groups as diverse as migratory vegetable pickers (Dunnette & Maetzold, 1955), army enlisted men (Beall & Blumenfeld, 1972), scientists (Smith, Albright, Glennon, & Owens, 1961), managers (Nash, 1966), nurses aides (Githens, Blumenfeld, & Hacker, 1967), clerical personnel (Fleishman & Berniger, 1960), and unskilled employees (Scott & Johnson, 1967). While the job titles have clearly varied considerably in the studies, all attempted (successfully) to predict a

personnel-oriented criterion such as turnover, accidents, supervisory ratings, and/or some index of production.

There has also been research in which this paradigm has been applied successfully toward the prediction of a criterion of financial credit risk (e.g., Buel, 1968) and, more recently, financial donation propensity (Blumenfeld & Sartain, 1974). For the most comprehensive single review of the literature on weighted application blanks, see England (1971).

Personnel Turnover

In an extensive review of the literature on forecasting personnel turnover, Schuh (1967) concluded that biodata has been a more successful predictor of this criterion than other potential predictors, e.g., achievement tests, personality measures, etc.

The accumulating experimental evidence offers considerable encouragement vis-a-vis the operating management problem suggested, i.e., the ability to forecast who among several individuals applying for enlistment has the greatest propensity for contract completion. Resource investment could be made so as to minimize planning and economic losses through personnel attrition.

Purpose(s)

There were two purposes of this research, i.e., content and method.

The content purpose was to predict completion or non-completion of military contract obligation based on biodata available at time of entry.

The method purpose was to evaluate the relative efficacies of three methods of assigning weights to biodata items.

Method

Data Collection

It should be noted that these data were obtained just before the elimination of the draft.

Criterion and participants. Criterion data (completion or non-completion of military contract obligation) and predictor data (previously collected at time of entry) were obtained for a sample of 400 enlisted men at a major Army installation in the Southeast. Of the sample, 200 men had completed their military contract obligation; and 200 men had not, i.e., had received an administrative discharge other than honorable or medical.

Predictor(s). The predictor data were 17 biodata items, i.e., standard demographic items, taken from the Form 20 (Enlisted Personnel Record). These biodata are available at time of (and prior to) entry.

Data Analysis

The data analysis consisted of four phases: (1) item analysis, (2) item weighting, (3) cross-validation, and (4) key (weighting method) efficacy comparison.

Item analysis. The two criterion groups were divided into primary and hold-out groups of 100 each. An item analysis was conducted in the primary group.

Item weighting. Three scoring keys (A, B, and C) were developed by assigning item weights in three different ways. These weighting procedures varied in relative sophistication from (A) arbitrary differential weights assigned by inspection to more sophisticated (B) differential weights and (C) unit weights.

The least sophisticated technique (A) was an inspection of the item analysis data, with differential weights being arbitrarily (based on a preconceived simple scheme) assigned as a function of the magnitude of the percentage difference observed. The more sophisticated differential weights (B) and unit weights (C) were what England (1971) terms net and assigned weights, respectively.

Cross-validation. The three weighting schemes (keys) were then applied to the records of the hold-out groups of 100 "completes" and 100 "non-completes." These cross-validations were quantified in terms of point biserial correlation.

Key efficacy comparison. Comparisons were then made among the obtained cross-validities.

Results

Based on the cross-validation data (N=200), the intercorrelations (product moment) among the three keys were: AB, .90; AC, .86; and BC, .84. The cross-validities (point biserial) of the three keys were: A, .62; B, .64; and C, .58.

Correlations of .14 are statistically significant beyond the .05 level; correlations of .18 are statistically significant beyond the .01 level.

Discussion

There were two purposes of this research, i.e., content and method.

Content Purpose

Regarding the first purpose of the study, i.e., to predict completion or non-completion of military contract obligation based on biodata available at time of entry, little discussion seems necessary. All three keys held up in cross-validation, i.e., achieved the conventional levels of statistical significance. While none approached perfection, all were within reasonable, practical, functional, magnitude expectation, possibly even greater than one might expect.

It thus appears possible to forecast, i.e., longitudinally over an approximate three-year period of time, at a better than chance level, the contract completion/non-completion behavior of this group of enlisted men based on biodata available at time of entry. This potential improvement in "batting average" should, or could, be of operational utility to management, particularly recruiters. That is, a (verbal) behavioral profile of the "kind" of individual the organization might best seek should aid management in recruiting and selection decisions.

Method Purpose

Regarding the second purpose of this research, i.e., to evaluate the relative efficacies of three methods of assigning weights to biodata items, the data are not quite so definitive, although a hierarchy of cross-validities did emerge. That is, while all three keys were cross-valid, B (.64) was greater than A (.62) ($p > .05$), and A (.62) was greater than C (.58) ($p > .05$).

In reaching a decision as to the relative efficacies of these three keys, several factors may be considered, e.g., their magnitude vis-a-vis zero, the magnitude of the difference between the key cross-validities, their ease of development, and their subsequent ease of operational scoring. Given the option of computerized item analysis, item weighting, and subsequent operational scoring, the "ease" considerations may be moot and academic, leaving one with the conclusion that key B, "sophisticated" differential (net), is to be recommended, statistical and practical significance considerations aside.

Without the option of computerization, however, the choice is more difficult and of a "trade-off" nature. That is do we select the greatest validity? (Key B), or ease of development? (Key A), or ease of operational scoring? (Key C). The conclusion is not so obvious, and would no doubt depend on local circumstances.

What is striking, however, is that the "quick-and-dirty inspection" differential key (A) and the somewhat more sophisticated differential (net) key (B) were both numerically superior (in varying degree) to the somewhat more sophisticated unit (assigned) key (C). This is somewhat at odds with the general finding in the literature that tends to suggest that there is no particular premium to be obtained by using differential weights. (Much of that work, however, was done prior to the real option of computerization.) In these data, however, regardless of the relative sophistication of the generation of the differential weights, differential weights appeared to be more efficient (at least numerically) than unit weights.

That the different weighting schemes were tapping essentially similar variance is reflected in the intercorrelation data, i.e., .90, .86, and .84. These intercorrelation data also tend to only partially bear out McCornack's caveat regarding the inference of validity merely as a function of the relationship between keys (1956). That is, keys having common variance do not necessarily relate in the same magnitude to a third, dependent, variable.

Data Limitations

The extent to which these data are peculiar to the 400 participants in this study is not known. Also, it is possible that the relative and/or absolute efficacies of these keys (or any others developed) might be affected by changes in world events, management policies, and/or other impinging incidents. Recall that these data were collected just before the elimination of the draft. Therefore, continued research seems called for.

Conclusions

The purposes of this research were to (1) predict completion or non-completion of military contract obligation based on biodata available at time

of entry, and (2) evaluate the relative efficacies of three methods of assigning weights to biodata items.

The data indicated that it is possible to predict completion or non-completion of military contract obligation based on biodata available at time of entry.

The data regarding the relative efficacies of the three methods of assigning weights to items were not quite so definitive. Although the obtained hierarchy of cross-validities does suggest differential weights to be superior (at least numerically) to unit weights, the data also suggest that the premium for sophistication in assigning weights not to be so great as one might expect.

Data limitations were indicated; and recommendation for continued research was made.

References

Beall, J. A., & Blumenfeld, W. S. Development of a psychometric procedure to forecast private motor vehicle accidents within a military sample. Proceedings of the 80th Annual Convention of the American Psychological Association, 1972, 7, 627-628. (Summary)

Blumenfeld, W. S. I am never startled by a fish. American Psychological Association Monitor, 1972, 3(9-10), 3, 14.

Blumenfeld, W. S. Application application. Atlanta Economic Review, 1973, 23(3), 8-13.

Blumenfeld, W. S., Godbey, W. L., & Blumenfeld, J. C. Application of a new psychometric method to an old tried and untrue experimental design to improve the validity of a tailor-made scoring key: or, Just Joshin'. Atlanta Economic Review, 1973, 23(3), 14-16.

Blumenfeld, W. S., & Sartain, P. L. Predicting alumni financial donation. Journal of Applied Psychology, 1974, 59, 522-523.

Buel, W. D. Evaluating mortgage loan risk. Journal of Applied Psychology, 1968, 52, 399-402.

Cureton, E. E. Reliability, validity, and baloney. Educational and Psychological Measurement, 1950, 10, 94-96.

Dunnette, M. D., & Kirchner, W. K. Validities, vectors, and verities. Journal of Applied Psychology, 1962, 46, 296-299.

Dunnette, M. D., & Maetzold, J. Use of a weighted application blank in hiring seasonal employees. Journal of Applied Psychology, 1955, 39, 308-310.

England, G. W. Development and use of weighted application blanks. Bulletin 55, Industrial Relations Center, University of Minnesota, 1971.

Fleishman, E.A., & Berniger, J. One way to reduce office turnover. Personnel, 1960, 37(3), 63-69.

Githens, W. H., Blumenfeld, W. S., & Hacker, R. L. Modern business practice: How hospital can improve aide tenure. Modern Hospital, 1967, 108(1), 20-22.

Kurtz, A. K. A research test for the Rorschach test. Personnel Psychology, 1948, 1, 41-51.

McCornack, R. L. A criticism of studies comparing item weighting methods. Journal of Applied Psychology, 1956, 40, 343-344.

Nash, A. N. Development of a SVIB key for selecting managers. Journal of Applied Psychology, 1956, 50, 250-254.

Schuh, A. The predictability of employee tenure: A review of the literature. Personnel Psychology, 1967, 20, 133-152.

Scott, R. D., & Johnson, R. W. Use of the weighted application blank in selecting unskilled employees. Journal of Applied Psychology, 1967, 51, 393-395.

Smith, W. J., Albright, L. E., Glennon, J. R., & Owens, W. A. The prediction of research competence and creativity from personal history. Journal of Applied Psychology, 1961, 45, 59-62.

PREDICTION OF TEST OUTLINE WEIGHTS FROM OCCUPATIONAL SURVEY DATA

David S. Vaughan, Capt., USAF
USAF Occupational Measurement Center

The United States Air Force Occupational Measurement Center, where this author is assigned, has a dual mission. The first mission is to construct Specialty Knowledge Tests and Promotion Fitness Examinations in support of the Weighted Airman Promotion System. The second mission is to conduct occupational surveys of various Air Force job specialties. As might be guessed, these two missions are closely related in many ways. The topic of this paper is one aspect of the relationship between occupational surveys and test construction. In particular, this paper is concerned with use of occupational survey results in determining the content of Specialty Knowledge Tests.

Air Force Occupational Surveys

At the USAF Occupational Measurement Center, the job content of various specialties is analyzed by an inventory method. This inventory method has been described in detail elsewhere (Morsh & Archer, 1967), but will be summarized here. The first step is development of a list of all duties and tasks potentially performed by an incumbent of a particular job specialty. A survey instrument is prepared which contains this task list, as well as background questions to obtain demographic information, work experience, and related data. The survey instrument is administered to all or a large number of the incumbents in the specialty to be surveyed. Incumbents indicate which of the tasks they perform in their present jobs, and rate, on a relative

scale, the amount of time spent on those tasks performed. Analysis of the survey results can provide data concerning the percent of members of a specified group who perform certain tasks or groups of tasks, and the percent of total time spent, on the average, on a task or group of tasks. The task list for a job specialty also provides a basis for collection of additional job factors. Currently, data on the relative difficulty of the tasks are routinely gathered.

Air Force Specialty Knowledge Tests

In the Air Force, promotion to certain enlisted grades is determined, in part, by performance on Specialty Knowledge Tests. These tests are designed to measure an airman's knowledge at a particular skill level in a particular Air Force job specialty (USAF Occupational Measurement Center, Note 3).

Air Force Specialty Knowledge Test construction is accomplished jointly by a test psychologist, who insures uniformity of style and clarity of item content, and a group of senior noncommissioned officers who serve as subject-matter specialists. These subject-matter specialists are brought to the Center on a short-term temporary duty basis from various operational locations, and they are able to provide an up-to-date picture of their job specialty from their several different perspectives.

A critical step in the test construction procedure is development and weighting of a test outline. This outline serves as a guide for the content of the individual items to be written. Team members are first asked to divide their specialty into broad functions. This list of functions is then

subdivided as appropriate and put into outline form. A sample outline portion is presented in Figure 1.

Insert Figure 1 about here

As a test outline emerges, subject-matter specialists are asked to assign percentage weights to each outline area. These percentage weights determine the number of test items to be written on each area of the specialty. The subject-matter specialists are instructed to base their outline weight assignments on the following factors: (1) criticality of the duties involved, (2) percentage of airmen in the specialty that perform related tasks, (3) percentage of time that the average airman in the specialty spends on related tasks, and (4) length of training time needed to acquire the skill, or "task difficulty". The purpose of outline weight assignments is to insure that the most "important" topics receive the largest number of test questions, and so on (USAF Occupational Measurement Center, Note 4).

Occupational Survey Data Test Construction

Data on several of the factors considered in the test outline weighting process are obtained by an occupational survey. Therefore, occupational survey results have an immediate application in the test construction process for outline weighting. In a paper presented at the 1974 Military Testing Association annual conference at Oklahoma City, Ms. Nancy Lewis (1974) talked about procedures currently followed at the Center for using occupational survey data in test outline weighting. If the large mass of data which is

obtained in an occupational survey is just given to the subject-matter specialists, the data are very difficult for them to use. The subject-matter specialists have difficulty just finding the relevent information among the huge quantity of data available, and they find interpretation of the data to be difficult when they can locate it. Therefore, a standard test construction computer product is prepared. In the abbreviated test construction product, data are provided for each of the three skill-level groupings of airmen for whom tests are to be constructed. Data are provided for each task on the survey task list and includes (1) the percent of airmen in the skill-level group who perform the task, (2) the average percentage of total time spent on the task by airmen in the group, and (3) the difficulty of the task.

Data presented in this form are much more accessible to the subject-matter specialists than such data were previously. However, subject-matter specialists still find that use of the data is difficult. The subject-matter specialists must translate data on three different variables for a list of over 500 tasks to numbers of items in outline areas which are much broader than specific tasks. Based on personal observation of the author, as well as discussion with other Center personnel, it appears that subject-matter specialists only use occupational survey data to resolve conflicts among themselves which cannot easily be resolved in other ways.

The special occupational survey computer product just described makes the survey data more accessible to subject-matter specialists. However, as was pointed out above, these survey data are still not systematically or heavily used. Two experiments will be discussed which illustrate a methodology

for converting occupational survey data into a form which might be more easily used by subject-matter specialists. As will be seen, this methodology has implications for content validation of tests. The experiments will also shed light on the judgmental process used by subject-matter specialists to assign outline weights.

Experiment 1

Purpose

The first experiment to be discussed was conducted by Karl Hickerson, myself, and other members of the USAF Occupational Measurement Center. The experiment is described in great detail elsewhere (Hickerson and Vaughan, Note 2). The purpose of this experiment was to examine the relationship between test outline weight assignments and several other variables which should be or might be important in those assignments.

Method

Two Air Force job specialties were selected for this study. These were 303X2, Aircraft Control Warning Repairman/Technician, and 403X0, Biomedical Equipment Maintenance Repairman/Technician. A total of 10 subject-matter specialists participated in this study, five from each specialty. These subject-matter specialists all held the grade of master sergeant (E-7) or above, and were selected for participation in test revision projects on the basis of their judged technical expertise.

A list of the variables used in this study is contained in Table 1.

Insert Table 1 about here

The criterion variable in all analyses was test outline weights, which were taken from outlines currently serving test development projects in the two job specialties. Three predictors, percent members performing, percent time spent, and task difficulty, were obtained from occupational survey results. The number of tasks associated with each outline area was determined from judgments of subject-matter specialists. Additional predictors were obtained by having subject-matter specialists rate test outline areas. These variables included additional job content factors--probability of emergency performance, and amount of detremental effect to poor performance; two factors relating to constraints in test writing--ease of writing and ease or reference test questions; and two other factors--importance for becoming fully qualified, and the completeness with which the occupational survey tasks covered outline areas.

The first step of the procedure involved having subject-matter specialists assign tasks from the occupational survey task list to various outline areas on the 7-skill-level outline in their particular job specialties. A task could be assigned to one outline area, several areas, or no area. The task assignments were done independently by each subject-matter specialist. Then the subject-matter specialists independently rated each outline area on each of the six rating variables. All task assignments and ratings were obtained for alpha-level outline areas. Outline area values for the occupational survey variables were obtained by summing across all tasks assigned to the areas. The percent members performing and percent time

spent variables from the occupational surveys were based on all airmen who responded to the surveys.

For the data analysis, a subject-matter specialist by outline area combination constituted one observation. The data included 200 such observations at the alpha outline area. Values for the Roman-level outline areas were obtained by summing across values for all alpha-level areas in each Roman-level area.

Results

Regression analyses were performed to determine the relationship between the predictors and outline weight. These analyses were performed separately for the alpha-level and Roman-level outline weights. Squared product-moment correlations between first-order predictor variables and outline weights at the two levels are presented in Table 2. As may be seen, correlations between outline weights and the predictors used in this study

---

Insert Table 2 about here

---

were generally very high at the alpha level and extremely high at the Roman-level. The increase in correlations between the alpha and the Roman-levels is not surprising, since the Roman-level outline weights had a much larger variance than the alpha-level outline weights (the alpha weight standard deviation was 3.13, while the Roman standard deviation was 13.73). Percent members performing, percent time spent, and number of tasks in the outline areas were less positively related than were any of the other factors.

Stepwise multiple regression analyses were performed to determine if outline weights could be predicted more accurately by a combination of the predictors than by each predictor taken individually. At the alpha outline level, squared and interaction variables were constructed from the first-order predictors and included in the regression analysis. $\underline{R}^2$ values for several of these multiple regression models are presented in Table 3.

Insert Table 3 about here

It may be seen from Table 3 that the squared and interaction predictors accounted for little variance beyond the simple linear set. Several of the predictors when taken alone, moreover, performed practically as well as three- and four-predictor equations in accounting for test outline weights. Because at the Roman outline level the first-order correlations were so high and because so little additional predictive ability was gained at the alpha-level from squared and interaction predictors, no such additional predictors were included in the Roman-level regression analysis. $\underline{R}^2$ values for selected regression equations at the Roman-level are presented in Table 4. It may be

---

Insert Table 4 about here

---

seen that little increase in $\underline{R}^2$ was obtained from inclusion of several predictors in these regression equations.

Difficulty and several of the predictors which were obtained by subject-matter specialists' ratings would serve individually as good predictors of outline weights at either outline level. Of these, difficulty is probably the best. This is suggested because the other variables may be very subject to criterion contamination. Subject-matter specialists had test outlines, included the weights, in front of them when they provided the ratings; seeing the outline weights may have influenced their ratings. Percent members performing and percent time spent had little relationship to test outline weights at either level. A possible explanation is that in this study, the occupational survey variables were based on all survey respondents rather than just those at the skill levels of the outlines. Another explanation is that the various outline areas had little variance in these variables.

Experiment 2

Purpose

The second experiment was designed as a replication and extension of the experiment described above. This experiment is being carried out by Captain Jim Hogan, myself, and other members of the USAF Occupational Measurement Center. Since data analysis has not yet been completed, only some preliminary results will be presented here. In this experiment, as in the previous one, regression models were examined for predicting test outline weights from occupational survey data aggregated for outline areas. This experiment is an extension of the first in that (a) two different job specialties were examined, and (b) prediction of test outline weights was tested at the three- or apprentice skill level and the five- or specialist

Skill level as well as at the technician skill level. This experiment, then, examined the generalizability of the previous results to additional skill levels and job specialties.

Method

The two job specialties selected for this experiment were 113X0A/C, Flight Engineer Specialist/Technician, and 902X0, Medical Service Specialist/Technician. In this experiment, as in the last one, subject-matter specialists assigned tasks from the appropriate occupational survey task list to the various skill levels. In this study, subject-matter specialists in each specialty worked as a group in task assignments rather than independently. Subject-matter specialists were not asked to rate outline areas for this study. Therefore, the following predictor variables were used to predict test outline weights: (a) average percent members performing for tasks in an outline area, (b) sum of percent members performing for tasks in an outline area, (c) average percent time spent, (d) sum of percent time spent, (e) average task difficulty, (f) sum of task difficulties and (g) number of tasks in an outline area. These variables are listed in Table 5. In

---

Insert Table 5 about here

---

contrast to the previous experiment, the percent time spent and percent members performing variables were based only on airmen at the particular skill level of the test outline, rather than on all airmen in the job specialty.

Results

Stepwise multiple regression analyses were performed at the alpha outline level for each skill level in each specialty. The criterion for all analyses was test outline weight, and predictors were those variables listed above. No squared or interaction predictors were used. First-order correlations between outline weight and each of the predictors are presented, at the alpha outline level, in Table 6. In some cases, the multiple regression analyses obtain $\underline{R}^2$ values that were substantially larger than the first-order correlations. In these cases, though, the increase

---

Insert Table 6 about here

---

was obtained mainly from use of predictors which correlated negatively with test outline weights and which had negative weights in their regression equations. From a practical standpoint, these negatively-weighted predictors do not seem useful. For example, it does not seem reasonable to give less weight to an outline area that has a higher percentage of members performing tasks in it. The most consistent result seems to be that the number of tasks assigned to outline areas performed well for prediction of test outline weights.

Discussion

Comparison of the two Experiments

The results of both experiments indicate that test outline weights can be predicted with some accuracy from occupational survey data. However,

the results of the two experiments seem to differ a great deal in the equation which should be used for making such predictions. These different results may be due to differences in procedure followed in the experiments. In the first, analyses were performed for both job specialties combined, and differences among the subject-matter specialists in their task assignments were taken into account. In the second, subject-matter specialists jointly assigned the tasks to outline areas in their specialties. Furthermore, separate analyses were conducted for each skill level in each specialty. Another explanation for the differences between the results of the two experiments is in differences between the job specialities and skill levels studied. As was mentioned above, data analysis for the second experiment is not yet complete. One goal of further data analysis in the second experiment to resolve the cause(s) for differences between results of the two experiments.

Content Validation

As an outgrowth of the research described above, a procedure could be implemented whereby subject-matter specialists would be provided test outlines with weights which are derived directly from occupational survey data. These weights would be obtained using equations designed to predict outline-weighting judgements of subject-matter specialists. The two experiments described illustrate a procedure for obtaining these prediction equations. Under such a system, occupational survey data would be presented in the form of numbers of test item to be written; occupational survey data in this form would be extremely accessible to the subject-matter specialists. Such a procedure for test outline weighting would also contribute to insuring the validity of tests.

Three main approaches to test validation are widely accepted. These are criterion validation, construct validation, and content validation (APA, 1974). Criterion or construct validation strategies do not seem practical for situations such as that at the present author's organization, where a large number of tests must be validated in a short period of time. Therefore, content validation is probably the most practical strategy for situations such as this. A number of different requirements are essential to successful content validation. Of these, perhaps the key requirement is the conversion of job analysis data into test content (EEOCC, Note 2). Conventionally, this requirement has been met through complete reliance on subject-matter specialists to insure that job characteristics are reflected in test content. A procedure for deriving test outline weights directly from occupational survey data, such as might be developed from the research discussed above, would be a more systematic, objective and reliable method for converting job content to test content.

In the research discussed above, the outline weight criterion of job importance was based on subjective judgements of subject-matter specialists, and the outline weight prediction systems were obtained to duplicate their judgments. Research indicates that, in general, statistical systems for prediction of judgmental criteria are likely to be more valid with respect to objective "ultimate" criteria than are the original judgments. (Dawes and Corrigan, 1974). If this finding can be applied to the present situation, one might conclude that outline weights assigned by a system which predicts judgments of subject-matter specialists will be better measures of the "ideal" outline weights than are the actual subject-matter specialists' assigned weights.

Unsolved problems

Several problems exist which should be resolved through future research before a test outline weighting system such as that proposed here is implemented. The first involves the procedure used for assigning tasks to test outline areas. Certain ambiguities arose during the task-assigning process. Some outline areas cover general principles or theory, rather than specific jobs. Such theory outline areas might be viewed as being related to virtually all of the tasks or none of the tasks. In either case, some distortion arises. One solution might be to have subject-matter specialists construct their outlines by sorting tasks into groups. This type of procedure could insure that every task belongs to exactly one outline area. However, this procedure would tend to make the association of test items and outline areas ambiguous.

The other main unsolved problem of the methodology proposed here for automated test outline weighting concerns how to combine occupational survey data to obtain the predicted outline weights. Results of the two experiments discussed above suggest contradictory methods for computing test outline weights from survey data. The contradictory results may be due to methodological differences between the two experiments or may reflect real differences among the various job specialties and skill levels examined. Further analysis of data from the second experiment may provide answers to this problem. Ideally, one equation could be used to compute outline weights for all Specialty Knowledge Tests. However, research may show that

separate equations are necessary for various groups of job specialties or individual specialties, or even different skill levels in each specialty.

Summary

In summary, a methodology was described which could be used for translating occupational survey data directly into test outline weights, which determine the numbers of test items to be written on various topics for Specialty Knowledge Tests. This methodology involves two main steps. First, subject-matter specialists assign tasks from an occupational survey inventory to various areas in a test outline. Then values for the occupational survey variables, including percent members performing, percent time spent and difficulty, are computed for each outline area. Multiple regression methods are used to arrive at an equation for combining the occupational survey data to obtain test outline weights. This equation can be used to convert occupational survey data directly into test outline weights for future test construction projects.

Results for the two experiments described here indicate that test outline weights can be predicted with some accuracy from occupational survey data using the described procedure. However, two unsolved problems remain. First, a better procedure for associating tasks from an occupational survey inventory with test outline areas is needed. Second, the necessity of different equations for converting survey data into test outline weights for different job specialties should be resolved.

Based on general findings concerning statistical prediction and modeling of human judgment from objective data (Dawes and Corrigan, 1974), one might conclude the methodology suggested here would produce "better" outline weights than do the subject-matter specialists. Such automatic weighting of outlines would contribute to the content validity of tests constructed from these outlines. It should be pointed out that the role of the subject-matter specialists would remain important. Besides performing assignment of tasks to outline areas, subject-matter specialists would be responsible for actual writing and verifying of test items and insuring that the actual content of the items matches the content of the outline areas.

Reference Notes

1. Equal Employment Opportunity Coordinating Council. Uniform Guidelines on Employment Selection Procedures (3rd draft). September 24, 1975.
2. Hickerson, Karl A., & Vaughan, David S. Content Validation by Computing Test Outline Weights from Job Analysis Data, unpublished manuscript, USAF Occupational Measurement Center, 1976.
3. USAF Occupational Measurement Center. Handbook for Construction of the SKT and Associated Tests. Lackland AFB, TX.: author, June 1, 1975.
4. USAF Occupational Measurement Center. Subject-Matter Specialist Hand Book for Construction of the SKT and Associated Tests. Lackland AFB, TX.: author, January 1, 1976.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. Standards for Educational and Psychological Tests. Washington: American Psychological Association, 1974.

Dawes, R. M., & Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Lewis, N. Use of occupational survey data in developing outlines for Specialty Knowledge Tests. Proceedings of the 16th Annual Convention of the Military Testing Association, 1974, 178-183.

Morsh, J. E., & Archer, W. B. Procedural Guide for Conducting Occupational Surveys in the United States Air Force (PRL-TR-67-11). Lackland AFB, TX.: Personnel Research Laboratory, Aerospace Medical Division, September, 1967. (NTIS No. AD-664 036).

Acknowledgments

The author wishes to thank Dr. Karl Hickerson and Capt. Jim Hogan, who collaborated on the research described here. In addition, thanks are due to 2Lt. Richard Holder, 2Lt. Scott Gregoire, A1c Bill Fogarasi, A1c Olaf Hallerud, and Amn. Linda Jones, who provided able assistance in data gathering and analysis, and all other members of the USAF Occupational Measurement Center who made the research possible.

TABLE 1

Experiment 1: Source and Definition of First-Order Variables

Variable	Source	Definition
Test Outline Weight[a]	Master outlines constructed for each test development project	Number of items to be included on test
Percent Performing	Occupational survey	Average percent members performing tasks within task groups (content areas)
Percent Time Spent	Occupational survey	Sum of percent time spent within task groups
Task Difficulty	Occupational survey	Average task difficulty values of tasks within task groups
Probability of Emergency Performance	Subject-matter specialist ratings	Probability that one or more tasks in group would be performed under emergency conditions
Detrimental Effect	Subject-matter specialist ratings	Amount of detrimental effect if the associated tasks in group were inadequately performed
Ease of Test-Item	Subject-matter specialist ratings	Ease of writing test items in the content area
Ease of Referencing	Subject-matter specialist ratings	Ease of referencing questions in the content area
Importance	Subject-matter specialist ratings	Importance of content area to being fully qualified in the job specialty
Completeness of Tasks	Subject-matter specialist ratings	Representativeness (completeness) of tasks assigned to each content area
Number of Tasks per Outline Area	Subject-matter specialist assignments	Number of tasks assigned to each task group

[a]This variable was the criterion for all regression analyses.

Table 2

Experiment 1: Squared First-Order Correlations Between Criterion and First-Order Predictors

Predictor	Squared Correlation	
	Alpha Level	Roman Level
Percent Performing	.05	.28
Percent Time Spent	.07	.12
Task Difficulty	.59	.92
Probability of Emergency Performance	.46	.92
Detrimental Effect	.59	.96
Ease of Item Writing	.27	.69
Ease of Referencing	.27	.69
Importance	.64	.96
Completeness of Tasks	.55	.92
Number of Tasks per Outline Area	.12	.14

Note. Squared correlations are with outline weights.

Table 3

Experiment 1: $\underline{R}^2$ Values for Selected Models at the Alpha Level

Predictors	$\underline{R}^2$
All predictors (including squares and interactions)	.816
All linear predictors	.710
All linear predictors plus interactions of survey variables with number of tasks in outline area	.717
Linear survey variables (percent time spent, percent members performing, difficulty, and number or tasks)	.622
Linear rating scales	.669
Linear and squared rating scales	.719
Percent time spent	.132
Difficulty, number of tasks, and detrimental effect	.660
Difficulty and detrimental effect	.636
Difficulty, number of tasks, and emergency performance	.628
Difficulty and number of tasks	.601
Percent time spent^2, percent members performing X difficulty, ease of referencing, ease of referencing^2, number of tasks	.617
All linear predictors except ease of writing and ease or referencing	.709
All predictors (including squared and interaction) which do not include ease of writing, ease of referencing, importance, or completeness of survey	.739
All linear predictors except ease of writing, ease of referencing, or completeness of survey	.709

Table 4

R^2 Values for Selected Models at the Roman Level

Predictors	R^2
All predictors	.982
Survey predictors	.951
Rating predictors	.972
Difficulty, detrimental effect, importance	.975
Percent time spent, percent members performing, importance	.946
Percent time spent, percent members performing, importance, detrimental effect	.975
Percent members performing, difficulty, detrimental effect	.975
Detrimental effect, emergency performance	.959
Difficulty, number of tasks	.932
Percent time spent, difficulty	.933
Percent members performing, difficulty	.983
Percent members performing, difficulty, number of tasks, detrimental effect	.975
Percent members performing, detrimental effect	.964

Note. No nonlinear (square or interaction) predictors were tested at the Roman level.

TABLE 5

Experiment 2: Definition of First-Order Variables

Variable	Definition
Test Outline Weight[a]	Number of items to be included on test
Average Percent Performing	Average percent members performing tasks within task groups (outline areas)
Total Percent Performing	Sum of percent members performing tasks within task groups
Average Percent Time Spent	Average percent time spent on tasks within task groups
Total Percent Time Spent	Sum of percent time spent on tasks within task groups
Average Difficulty	Average difficulty of tasks within task groups
Total Difficulty	Sum difficulties for tasks within task groups
Number of Tasks	Number of tasks assigned to task groups

[a]This variable was the criterion for all regression analyses.

TABLE 6

Experiment 2: First-Order Correlations Between Criterion and First-Order Predictors

Job Specialty	Average Percent Members Performing	Total Percent Members Performing	Average % Time Spent	Total % Time Spent	Average Difficulty	Total Difficulty	Number of Tasks
11330A	- .21	.06	- .29	- .02	.53	.34	.23
11350A	.10	.71	- .02	.66	- .38	.80	.80
11370A	.02	.71	.02	.68	- .38	.80	.80
11370C	.23	.20	.17	.19	.21	.18	.18
90230	.35	.38	.19	.37	- .08	.51	.49
90250	.02	.69	- .20	.64	- .29	.65	.67
90270	- .46	.28	- .50	.24	- .09	.39	.39

459

CONTROLLED ITEM (TEST MATERIAL)

AFSC/PT NR & DATE	AFS	REVISION	SHREDOUT	PROJECT DATE	FOR FINAL COPY
31670 1 Jul 70	E-7 MISSILE SYSTEMS ANALYST TECHNICIAN (Shredouts A & B)	03	COMMON	2 Jan 70	

SUBJECT MATTER AREA	PER CENT PER AREA (100)	ITEMS REQUIRED	ITEMS IN FINAL TEST (140) (68)	ITEM NR IN FINAL TEST
I. ANALYZES FUNCTIONAL DATA COLLECTED FROM MISSILE TELEMETERING SYSTEMS	5	8	6	3,4,5,6,7,8
II. APPLIES INFORMATION OBTAINED DURING VARIOUS STAGES OF MISSILE DEVELOPMENT AND TESTING PROGRAMS	12		14	
A. Uses Launching Information	4	6	5	9,10,11,12,13
B. Uses Flight Data	4	6	5	14,15,16,17,18
C. Uses Automatic Homing	4	5	4	19,20,21,22
III. APPLIES MAINTENANCE MANAGEMENT PRINCIPLES	15		17	
A. Collects Maintenance Data	5	8	6	23,24,25,26,27,28
B. Schedules and Controls Work	5	8	6	29,30,31,32,33,34
C. Uses Supply Procedures	5	6	5	35,36,37,38,39
IV. SUPERVISES AND MANAGES MISSILE SYSTEMS ANALYSIS PERSONNEL	12		14	
A. Practices Missile Safety	3	4	3	40,41,42
B. Follows Security Regulations	4	6	5	43,44,45,46,47
C. Supervises Administrative Functions	5	8	6	48,49,50,51,52,53

NR OF NEW ITEMS 35 | NR OF VERBATIM ITEMS 6 | NR OF REVISED ITEMS 10 | PAGE 1 OF 5 PAGES

AF [illegible] 70 0-25 (PA) PREVIOUS EDITIONS ARE OBSOLETE. SPECIALTY KNOWLEDGE TEST OUTLINE

CONTROLLED ITEM (TEST MATERIAL)

460

516

Fig. 1. Specialty Knowledge Test Outline

517

SYMPOSIUM

COMPUTERIZED ADAPTIVE TESTING AND ITS POTENTIAL FOR MILITARY TESTING APPLICATIONS

A BRIEF OVERVIEW OF ADAPTIVE TESTING

James R. McBride

U. S. Army Research Institute for the Behavioral and Social Sciences

This symposium is intended as an exposition of some potential applications of adaptive testing to several military testing problems. My purpose as chairman is to provide you with a necessarily brief introduction to adaptive testing--what it is, what is needed to implement it, and why we should be interested in it.

"Adaptive" testing is just one of a number of terms used to describe a procedure whereby the test items that comprise an individual's test are selected during the test itself. Some terms used interchangeably with adaptive testing include tailored testing, branched testing, programmed testing, individualized testing, and others. We prefer the term adaptive because these tests adapt themselves to the examinee; different persons answer different items, with the items chosen sequentially to suit the individual examinee's performance.

Differential selection of test items may be accomplished in any number of ways, but the rationale is quite simple. Choose a more difficult item following each correct answer, an easier item following an incorrect one. Some methods of adaptive testing have been implemented in paper-and-pencil mode; for example, Lord's (1971) flexilevel adaptive test was designed specifically for paper-and-pencil administration. However, experience has shown that the instructions for paper-and-pencil adaptive tests are too complex for some examinees to follow successfully. A more satisfactory mode of administration is by means of a computer terminal or similar device. Thus, Weiss (1976) chose to administer adaptive tests at a CRT terminal; Bayroff, Ross and Fischl (1974) reported the Army's development of a computer-controlled slide projection terminal for adaptive testing; Waters, et al. (1976) designed and built a microprocessor terminal which directs the examinee through an adaptive sequence of test items read from a printed booklet.

Item Selection Strategies. Because adaptive tests are quite different from conventional tests in which all examinees must answer the same set of test items, adaptive testing poses some unique psychometric problems. One problem is how to choose the next item, from the pool of available items. The solution is an item selection strategy, a formalized rule for item choice.

There is no limit to the number of strategies possible. They run the gamut from very simple two-branch rules to rules based on the optimization of rather complex mathematical functions. Obviously, computerizing the item selection process facilitates the use of the mathematical optimization procedures.

Scoring Adaptive Tests. Since different examinees take sets of test items which may differ in number, difficulty, and discriminating power, the traditional number correct score will not suffice as an index of order of merit on most adaptive tests. Some scoring procedure is required which will consider not only how many items were answered correctly, but also which items were taken, and the pattern of right and wrong answers to those items. The scoring procedures must widely used in adaptive testing are based on various formulations of latent trait theory (e.g. Birnbaum, 1968, Lord, 1952, 1974; Rasch, 1960). All these formulations support statistical methods for locating examinees on a common scale, even though they responded to different sets of test items. These statistical methods all involve estimating the examinee's position on the scale, based on the pattern of his responses to the items taken.

Item Response Theory. The unique characteristics of adaptive tests--tailoring each test to the individual, scoring all tests on a common scale despite the different items constituting each test--make traditional test theory inadequate to deal with it. Latent trait theory, or item response theory (Lord, 1952, 1976), has been developed in time to provide an adequate theoretical basis for the development of adaptive testing. One might say that if item response theory didn't exist adaptive testing specialists would have had to invent it.

Item response theory is a general term for theoretical formulations which account for people's responses to test items in terms of their status on an underlying attribute. In ability (or achievement) testing, we say that the higher the attribute status, the larger is the probability of a correct response to any given item which measures the ability in question. Through appropriate scaling procedures, we can construct a response model for every such test item. This model, an item characteristic curve (i.c.c.), expresses the probability of a correct response as a mathematical function of scaled ability.

Every person can be characterized by his/her location on this ability scale. Every test item also has a location parameter (its threshold, or "difficulty") and perhaps a unique scale parameter (proportional to the slope of the i.c.c.), analogous to its discriminating power. Some items have also a lower asymptote, or guessing parameter.

Knowing which items a person has answered; the difficulty, discrimination and guessing parameters of those items; and whether the answers were correct or incorrect, permits us to use the statistical techniques of item response theory to estimate the examinee's ability. The resulting ability estimate is a "test score" of sorts, and as such it has an error component like any other observed score. Unlike classical test theory, item response theory makes no assumption that measurement errors are independent of "true score". That is just as well, because that central assumption of classical test theory is untenable anyway (Lumsden, 1976). Whether ability is defined as "true score," or as location on a latent continuum, errors of measurement vary systematically as a function of ability level, reflecting the discrepancy between examinee ability and item difficulty, the discriminating power of the test items, and the effects of guessing.

Item response theory permits us to evaluate something closely akin to the standard error of measurement as a function of underlying ability, if we know the test item parameters. This "something" is called the test information function (Birnbaum, 1968), and is inversely proportional to the standard error of estimating an examinee's location on the ability scale. If we were to plot the information function of a typical peaked conventional test (one whose items are all about equal in difficulty), we would observe that its test information function is likewise peaked--very high over a narrow range of ability, but diminishing in magnitude elsewhere. Such a test will discriminate well over a narrow interval of the ability range; it will not discriminate very well outside that interval. Let us refer to the ability level at which the test information function is highest, as the test "center".

If we were now to look at the information function of a "rectangular" conventional test--one whose item difficulties are uniformly distributed over a wide range--we would notice that the information function was fairly flat but low over a broad interval on the ability scale around the test center. This test would measure about equally well over a much wider range that the peaked test, but other things being equal, would not discriminate nearly as effectively as did the peaked test at its center.

Now, a test measures best (most precisely, if you will) where its information function is highest (and hence its standard error is lowest). We would like to have high measurement precision over the entire normal range of variation of the attribute we seek to measure. This is tantamount to a high, flat information function. In conventional testing, however, we have a dilemma. We can build a peaked test, which yields a peaked information function; or we can go to the other extreme and build a rectangular test, which has a low, flat information function. We cannot build a test with a high, flat information function unless we make it extremely long.

I like to think of this as a "bandwidth-fidelity" dilemma, with apologies to Cronbach (1960). The designer of a conventional test can construct it to have high "fidelity"--high precision, low measurement error--over a narrow range of ability; or to have broad "bandwidth"--equiprecision of measurement over a wide range of ability, at the expense of fidelity. In designing a conventional test there is a tradeoff between broad bandwidth and high fidelity; the designer cannot have both.

Herein resides the most attractive feature of adaptive tests from a psychometric point of view: Because the test is adaptive to the individual, the discrepancy between ability and item difficulty can be made both small and fairly constant across the ability range. The result is a flat information function which is also generally high. Adaptive tests--and only adaptive tests--are capable of accurate, equiprecise measurement over a wide ability range. This pays dividends in test reliability, and criterion validity, and in the general utility of the test for a broad range of measurement and decision applications.

A properly designed adaptive test will have higher reliability than a conventional test of the same length. As a corollary to that, an adaptive test can achieve a specified level of reliability in materially fewer items than a conventional test, thus permitting the measurement of additional dimensions in the time saved. Both improved reliability and additional measurements should result in an increment in predictive validity over that obtained using conventional tests.

In addition to the psychometric benefits accruing from the use of adaptive tests, there are psychological benefits to the examinees. Adaptive tests can have positive effects on the test-taking motivation of examinees (see, for example, Betz, 1976); and by tailoring test difficulty to examinee ability, they can both reduce the effect of guessing and make any remaining effect relatively constant across ability levels.

Summary

I have given a rather broad-brush treatment to adaptive testing in my remarks to this point. Hopefully that treatment has left you with a conception of what adaptive testing is, of the rudiments of the test theory that supports it, and of the significant psychometric and psychological advantages that can accrue when a well-designed adaptive test program is implemented in a mental measurement setting. The four principal papers in this symposium will deal in more detail with some methods used in conjunction with adaptive testing, and with a variety of areas of application of adaptive tests which are relevant to the needs and problems of test users in the military.

References

Bayroff, A. G.; Ross R. & Fischl, M. A. Development of a programed testing system. Technical Paper 259. Arlington, Virginia: Army Research Institute for the Behavioral and Social Sciences, December, 1974.

Betz, N. E. The effects of immediate knowledge of results on computer-administered tests of verbal ability. Unpublished doctoral dissertation, University of Minnesota, 1976.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.

Cronbach, L. J. Essentials of psychological testing. New York: Harper and Row, 1961.

Lord, F. M. A theory of test scores. Psychometric Monograph. No. 7, 1952.

Lord, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151.

Lord, F. M. Individualized testing and item characteristic curve theory. In Krantz, Atkinson, Luce, & Suppes, Contemporary developments in mathematical psychology. San Francisco: W. H. Freeman, 1974.

Lord, F. M. Applications of item response theory to practical testing problems. Paper presented at the convention of the American Psychological Association, Washington, D. C., September, 1976.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 251-280.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogische Institut, 1960.

Waters, B. K. Development of a microcomputer-controlled, interactive testing terminal. Lowry AFB, Colorado: Air Force Human Resources Laboratory, in preparation, 1976.

Weiss, D. J. Computerized Ability Testing, 1972-1975. Final Report. Minneapolis: Psychometric Methods Program, Department of Psychology, University of Minnesota, April, 1976.

Estimation of Latent Trait Status Using Adaptive Testing Procedures

by
James B. Sympson
University of Minnesota

During the last few years, latent trait theory has become increasingly important as a theoretical foundation for the practice of psychological and educational assessment. This has been partly due to shortcomings inherent in classical test theory (see the chapter by Lumsden in Annual Review of Psychology, 1976) and partly due to recent developments in testing practice. In particular, when "adaptive" or "individualized" testing is desired, latent trait theory provides a uniquely useful conceptual scheme for guiding test design and test scoring procedures.

Latent trait theories are characterized by a mathematical model that relates the probability of occurrence of a particular response class (e.g., a "correct" response) in the presence of a particular stimulus (e.g., a test item) to a person's position on one or more metric dimensions. The graph of the function that relates probability of a particular response class to a person's status on these dimension(s) can be referred to as a response-characteristic curve.

Both univariate and multivariate latent trait models have been proposed. The univariate models (e.g., Lord, 1952; Rasch, 1960; Birnbaum, 1968; Bock, 1972) assume that response probabilities are related to the relative positions of persons and stimuli on a single metric dimension. Multivariate models (e.g., Samejima, 1974; Christoffersson, 1975) allow for the possibility of several latent dimensions.

Latent Trait Theory and the Objectives of Measurement

Before proceeding, I would like to address certain issues that trouble many people when they first encounter latent trait theory. It is reasonable to ask, "Why should I bother with an approach to testing that involves inferred latent traits, when what I'm really interested in is (choose one: predicting some criterion accurately; content validity; criterion-referenced measurement)?" While I do not have time to present extensive supporting arguments here, the following comments provide the essence of my response to these questions.

First, the adoption of latent trait theory as a guide to test construction and scoring practices does not require a belief in the "existence" of unobservable traits that control human behavior. Empirically, it is sufficient to inquire whether peoples' responses

---

This research was supported by Contract N0014-76-C-0243, NR 150-382, with the Personnel and Training Research Programs, Office of Naval Research, David J. Weiss, Principal Investigator.

to test stimuli can be well predicted on the basis of such a model. The postulated dimensions of latent trait theory can be viewed as quantitative variables that we create by calibrating and scoring test items in a certain way. Needless to say, it is our hope that the variables we create will provide a convenient basis for designing testing procedures and will lead to increased predictive accuracy in scientific and practical applications.

In many situations tests are developed and applied with the sole intention of predicting performance in a criterion setting. The introduction of intervening variables (latent traits) might seem unnecessary when one is only interested in obtaining a high degree of relationship between test scores and criterion scores. I have two comments to make regarding this issue. First, estimates of latent trait status can themselves be viewed as a particular variety of test score. Such scores may or may not have higher predictive validity than more conventional test scores. This is an empirical question. However, even if predictive validity is not increased through the use of latent trait scores, it may still be advantageous to adopt a latent trait approach if the testing process can be made more efficient as a result (e.g., through adaptive testing procedures).

Second, it is my contention that test development for the purpose of criterion prediction is always based upon an implicit structural model. No one chooses items at random from all conceivable item domains. We try out items with certain kinds of content and never consider using other kinds of content. We also pay careful attention to generating items that have difficulty levels or endorsement rates (i.e., p-values) that are not too extreme in the population to be tested. This is done so that item-criterion correlations will not be unduly restricted. Such procedures suggest the existence of an implicit structural model.

Trying certain types of items, and not others, implies that certain types of inter-person differences exist and are related to criterion performance, while others are not. More generally, any conceptual scheme for classifying test items implies a corresponding set of response variables that can be generated when the items are administered. In selecting items for criterion prediction we indicate the type of response variable(s) we think to be related to the criterion.

A concern about item difficulties and endorsement rates implies that the probability of a given response to an item is a function of status on the relevant response variable(s). If such probabilities were not a function of status on the response variables, an item would have the same p-value in each and every population we could conceive of and there would be no need to match item difficulties to the population we plan to test.

A latent trait approach to test construction and scoring provides a formal vehicle for elaborating structural models and encourages the test developer to bring implicit structural assumptions out into the

open. Explicit structural models can serve to guide our test construction efforts and aid in the interpretation of empirical results.

With regard to content validity and criterion-referenced measurement, it is important to keep in mind that the testing situation never constitutes the entire domain of interest. Our implicit objective in pursuing content validity and in implementing criterion-referenced measurement is to make more accurate inferences about a person's potential for performance in a hypothetical task domain (Cronbach, 1971, p. 452; Glaser & Nitko, 1971, p. 653). This hypothetical task domain, though it is not observable in its entirety, is always carefully defined in terms of performance objectives or item content. Test items are generated that represent the domain, and responses to these items are used as a basis for making inferences about domain performance.

Some individuals protest such a view and argue that in criterion-referenced measurement the test stimuli *are* the criterion tasks of interest and that no further task domain is intended or implied. If this were in fact the case, we could develop a criterion-referenced test for, say, aircraft mechanics and have students remain in technical school for the entire duration of their enlistment. If the tasks in the test *are* the criterion tasks of interest, then performance on the test would constitute a sufficient basis for writing ERs (efficiency reports) and determining who should be promoted. There would be no need to have people actually go into the field and work on operational aircraft. The fact that criterion-referenced tests are being developed for use in field settings does not defuse this obviously absurd conclusion. Unless we include each and every relevant task completed by an aircraft mechanic during each and every work day in the "test", we are necessarily making inferences about a larger task domain from a sample of person-stimulus interactions drawn from the domain.

What is the nature of the hypothetical task domain in achievement testing? Given the emphasis on generating items that fulfill various content or process requirements and the emphasis on learning hierarchies (i.e., difficulty hierarchies) that permeates educational testing, I would suggest that such task domains can be described in terms of a multidimensional structural model. Whenever test stimuli can be clustered with regard to common content or process and then ordered within a cluster, there is a definite possibility that a latent trait approach to achievement testing will be useful.

In recent years the distinction between norm-referenced and criterion-referenced measurement has been widely discussed. An important fact to keep in mind is that this distinction properly applies to the type of information available from test scores, not to test content or the testing procedure itself (Hambleton & Novick, 1973, p. 162). This is an important fact because, as it turns out, estimates of latent trait status can provide information about both inter-person differences (norm-referenced interpretations) and intra-

person response probabilities for tasks drawn from a task domain (criterion-referenced interpretations).

An estimate of an individual's latent trait status can be converted to a centile rank or standard score relative to any norm group previously tested using the latent trait procedure. This same latent trait estimate, when considered in conjunction with the latent trait parameters of a test item (i.e., a task sample) that has been previously calibrated, allows us to generate the probability of occurrence of a given response class (e.g., a "correct" response) in the presence of the item. The fact that latent trait theory can provide both norm-referenced and criterion-referenced interpretations of test performance indicates that the current schism between psychological and educational testing may be narrowed considerably in the years to come.

Estimating Latent Trait Status

Hopefully, at this point I have convinced you that latent trait theory has a wide range of potential applications. If so, then procedures for estimating latent trait status should also be of interest to you. In the following discussion I will present four methods for obtaining estimates of latent trait status and will argue that the accuracy of such estimates can be improved through the use of adaptive testing procedures.

In my example, I will deal with a latent trait model in which only two response classes are considered, a *keyed* response and a *non-keyed* response, and where the probability of occurrence of each response class is a function of a single latent dimension. This model might be applicable to a test that has been constructed to maximize internal consistency (Nunnally, 1967, pp. 254-268) and in which items are scored dichotomously. The model would not be suitable for tests that involve a multidimensional item structure, but the principles of latent trait estimation that I will present are generalizable to such cases.

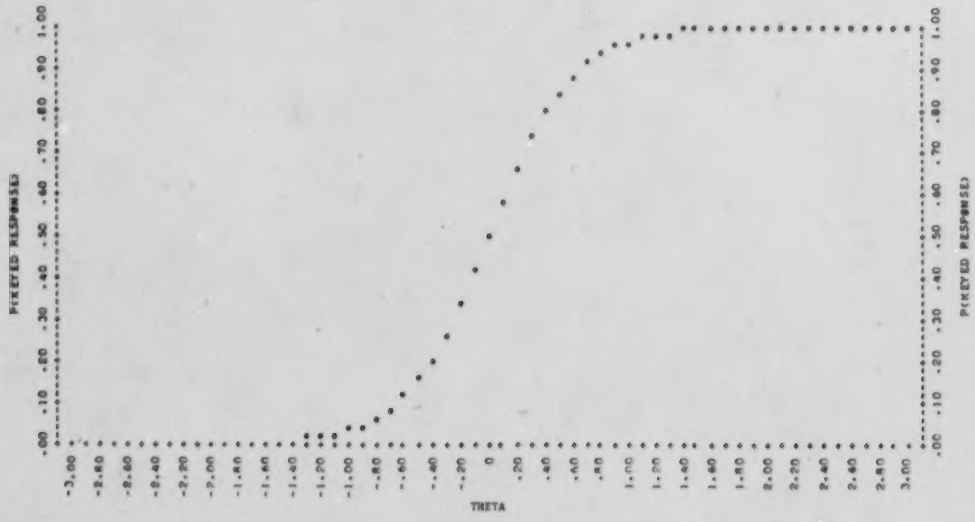
The three-parameter logistic model. The latent trait model that I will use in my example has been investigated extensively by Birnbaum (1968). This model is referred to as the *three-parameter logistic* model, and the function rule that relates probability of a keyed response to the parameters of the model is given in Equation 1.

$$P_g(\theta) = c_g + (1-c_g)[1+\exp(-1.7a_g(\theta-b_g))]^{-1} \qquad [1]$$

The quantity $P_g(\theta)$ is the probability of a keyed response to item g, with parameters a_g, b_g and c_g, by a person whose location on the latent trait continuum is given by the quantity θ (theta). The exponential operator (exp) indicates that the quantity in parentheses is an exponent of the constant $e \doteq 2.71828$.

Figure 1 shows a graph of the function $P_g(\theta)$ in the interval from θ=-3.00 to θ=+3.00 for an item having a_g=2.0, b_g=0.0, and c_g=.00. This graph was generated by evaluating $P_g(\theta)$ at 61 points along the theta continuum. The irregularities visible in Figure 1 result from rounding $P_g(\theta)$ to the nearest .02 for plotting purposes.

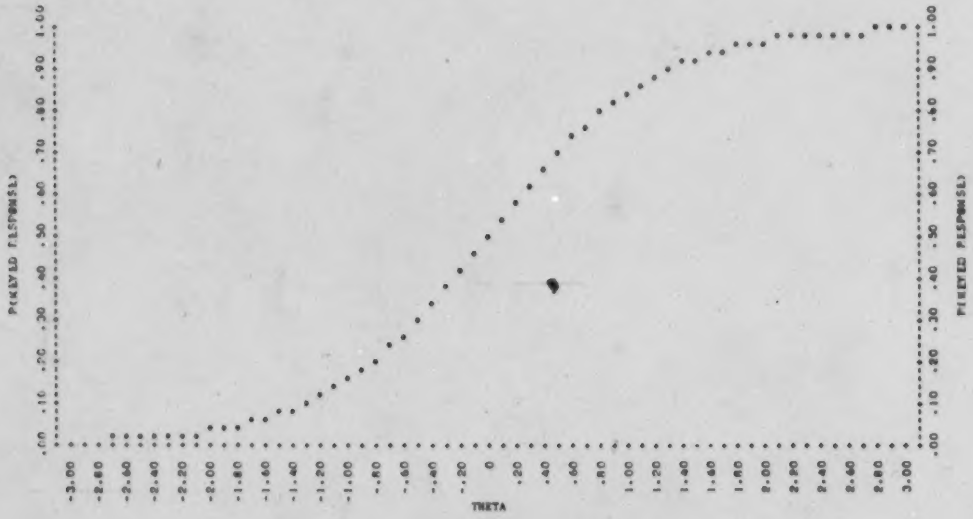
Figure 1
Response Characteristic Curve (a=2.0, b=0.0, c=.00)



The item parameter b_g is known as the item *difficulty* parameter. It is equal to the location on the latent trait continuum at which $P_g(\theta)$ is equal to $.5(1+c_g)$. The item parameter a_g is known as the item *discrimination* parameter. It is related to the slope of the response characteristic curve and in this model is equal to the reciprocal of the distance that one must move along the theta continuum in order to increase $P_g(\theta)$ from $.5(1+c_g)$ to approximately $(.8455(1-c_g))+c_g$. Since a_g=2.0 and c_g=.00 in Figure 1, the distance between the locations on the theta continuum at which $P_g(\theta)$=.5 and $P_g(\theta)\doteq.84$ is equal to $1/a_g$=.50 theta units. The item parameter c_g is the value of $P_g(\theta)$ when $\theta=-\infty$. It is the lower asymptote of $P_g(\theta)$ and is usually conceived of as the probability of a keyed response occurring "by chance" when $\theta=-\infty$.

Figure 2 shows a response characteristic curve for an item having a_g=1.0, b_g=0.0, and c_g=.00. The reduced value of a_g, relative to Figure 1, is reflected in the shallower slope of this graph and in

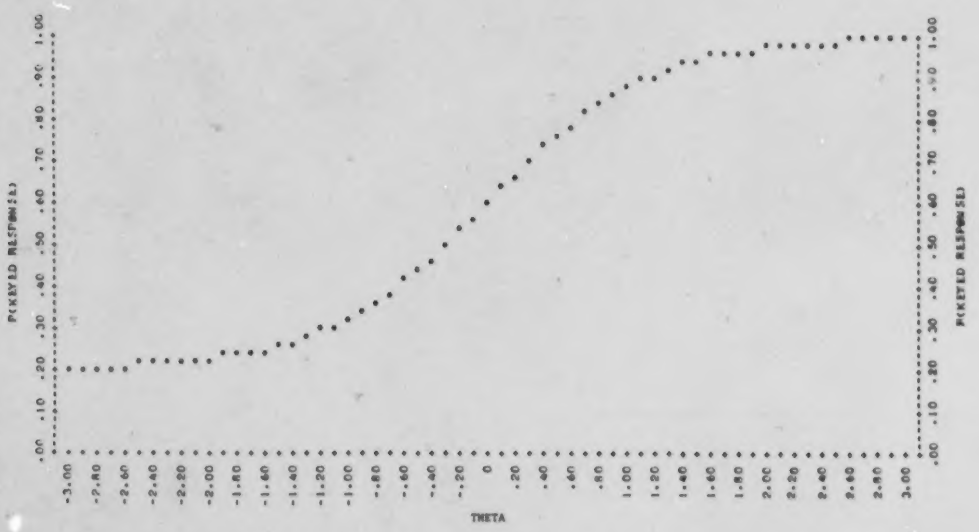
Figure 2
Response Characteristic Curve (a=1.0, b=0.0, c=.00)



the fact that the distance between the locations at which $P_g(\theta)$=.50 and $P_g(\theta)\doteq$.84 is now equal to $1/a_g$=1.00 theta units. A value of a_g in the vicinity of 1.0 is typical of many test items. Values of a_g below about .5 are indicative of "poor" items and values of a_g above 2.0, while desirable in many applications, are not common.

Figure 3 shows a response characteristic curve for an item having a_g=1.0, b_g=0.0, and c_g=.20. The value c_g=.20 might be applicable to a multiple-choice test item that has five response alternatives. In accord with the definitions given above, we see that b_g is equal to the location at which $P_g(\theta)$=.5(1+.2)=.60 and that a_g is equal to the reciprocal of the distance from the location at which $P_g(\theta)$=.60 to the location at which $P_g(\theta)\doteq$(.8455(1-.2))+.2$\doteq$.88. Note that one of the effects of a non-zero c_g is to reduce the slope of $P_g(\theta)$ at all points along the theta continuum.

Figure 3
Response Characteristic Curve (a=1.0, b=0.0, c=.20)



The concept of "information". Birnbaum (1968) has discussed the concept of "information" available in a test item. Birnbaum's *item information function* is given in Equation 2.

$$I(\theta, u_g) = [P'_g(\theta)]^2/[P_g(\theta)\ Q_g(\theta)] \qquad [2]$$

In this equation, u_g is the *item response variable*. It is equal to 1 when a keyed response is emitted and is equal to 0 otherwise. The quantity $Q_g(\theta)$ is equal to $1-P_g(\theta)$. The numerator of Equation 2 is the squared first derivative (i.e., the squared slope) of $P_g(\theta)$ at a fixed value of θ. The denominator is the variance of the item response variable, u_g, at a fixed value of θ. The quantity $I(\theta,u_g)$ is an index of how well we can discriminate people whose latent trait location equals θ from people at nearby latent trait locations on the basis of responses to item g.

In general, a steeper slope for $P_g(\theta)$ implies greater discriminating power. As we noted earlier, high values of a_g and low values of c_g increase the slope of $P_g(\theta)$ and, hence, the information available from an item. The variance of u_g approaches zero at latent

trait levels that are deviant from b_g and reaches its maximum value at the latent trait level where $P_g(\theta)=.5$. Figure 4 shows a graph of the function $I(\theta,u_g)$ in the interval from $\theta=-3.00$ to $\theta=+3.00$ for an item having $a_g=1.0$, $b_g=0.0$, and $c_g=.00$. This graph was generated by evaluating $I(\theta,u_g)$ at 61 points along the theta continuum and rounding the obtained values to the nearest .02.

Figure 4
Information Curve for a Single Item (a=1.0, b=0.0, c=.00)

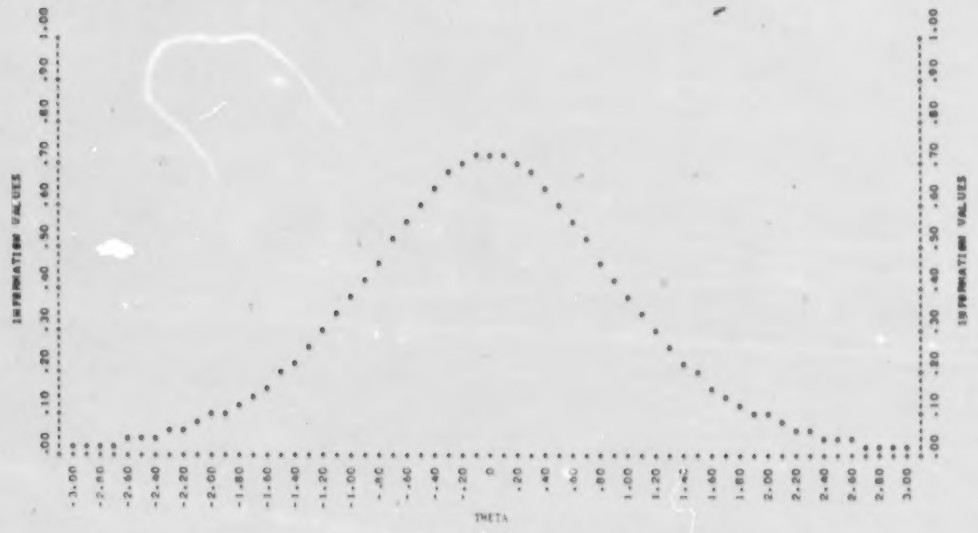


Figure 4 shows that an item provides maximum information in the region of the theta continuum where the item is located (i.e., near b_g) and relatively little information at levels far below or far above b_g. This result is consistent with intuitive impressions of item discriminating power. If, for example, we gave an ability test item that was suitable for third graders (i.e., $P_g(\theta)$ near .5 among third graders) to college students (in which group $P_g(\theta)\doteq 1.0$), we would find that all the college students answered correctly and no basis for discriminating among college students would exist. Note that in Figure 4 the information curve is symmetric about b_g and attains a maximum value of approximately .72.

Figure 5 shows an information curve for an item having $a_g=.85$, $b_g=0.0$, and $c_g=.00$. This curve, while still symmetric about b_g,

attains a lower maximum (approximately .52) and falls off more gradually on either side of b_g than the curve in Figure 4. In fact, the item represented in Figure 5 provides slightly more information than the item represented in Figure 4 in the interval below $\theta \doteq -1.40$ and in the interval above $\theta \doteq 1.40$. However, the gain in these regions is slight compared to the information loss in the interval $-1.40 \doteq \theta \doteq 1.40$.

Figure 5
Information Curve for a Single Item (a=.85, b=0.0, c=.00)

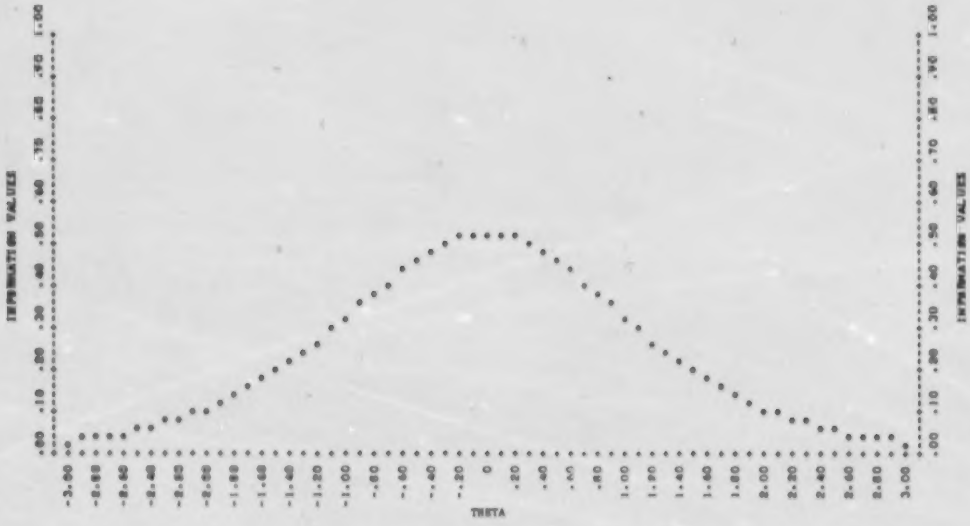
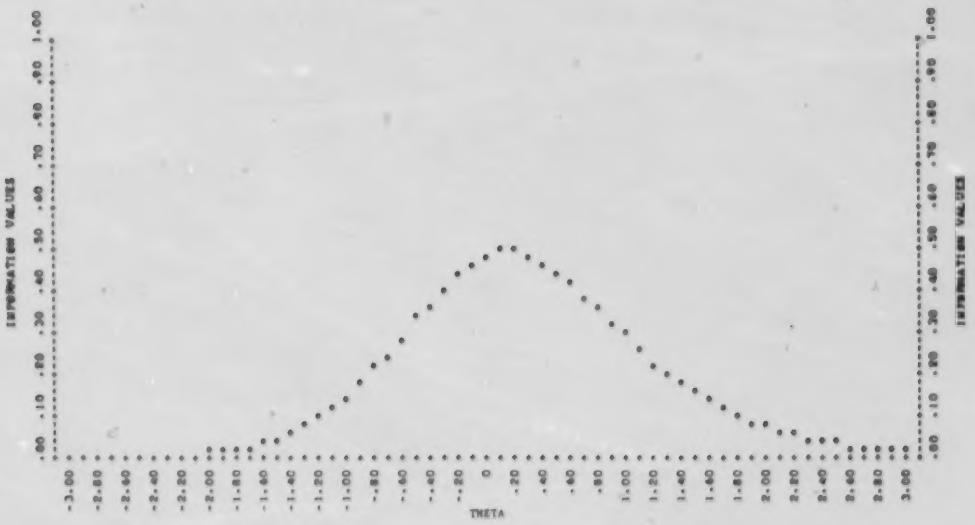


Figure 6 shows an information curve for an item having a_g=1.0, b_g=0.0, and c_g=.20. This curve is not symmetric about b_g. It attains its maximum value of about .50 near θ=.16. The curve falls off more rapidly on the left of θ=.16 than on the right. This reflects the fact that "chance" keyed responses are more prevalent among people located below b_g than among people located above b_g. Such "lucky" responses contribute error to the estimation of latent trait status and reduce the amount of information available. We may note that a_g=1.0, c_g=.20 produces an information curve that is lower than the curve in Figure 5. Thus, introducing the possibility of "lucky" keyed responses reduces the information available from

an item just as if we had selected an item with lower a_g, but with c_g=.00.

Figure 6
Information Curve for a Single Item (a=1.0, b=0.0, c=.20)



Sequential estimation in an adaptive test. In order to demonstrate the sequential estimation of latent trait status in an adaptive test, a computer program was used to simulate the test responses of a person whose latent trait location is θ=+1.0. A total of 20 items having a_g=1.0 and c_g=.20 was administered. The item b_g values changed from item to item as a function of responses generated during the simulated test. Table 1 summarizes the results of this 20-item test.

The first column in Table 1 contains item numbers in the 20-item series (g=1,2,...,20). The second column contains the b_g values of the items administered. The difficulty of the first item was b_1=0 because this value approximates the mean latent trait score in any population of persons that is sampled to parameterize a set of test items. (An exception to this may be found in Wright & Panchapakesan's (1969) implementation of the Rasch model. They scale the latent trait metric such that the mean of the b_g estimates is zero and the mean θ estimate among persons is, in general, other than zero.) Following the first item, b_g values either increase or decrease (in

accordance with a procedure that I will outline in a moment) depending on whether a keyed or non-keyed response was generated. The item response variable u_g is shown in the third column of Table 1.

Table 1
Sequential Estimation of Latent Trait Status
in a 20-Item Adaptive Test

Item No.	Diff.	Resp.	MAXL. Est.	W.B.L. Est.	SBAYES Est.	OBAYES Est.
1	0	1	5.49	1.61	.38	.38
2	1.00	0	.36	-.85	.05	.04
3	0	1	.67	.18	.32	.31
4	.18	1	.89	.82	.53	.54
5	.82	1	1.16	1.25	.75	.78
6	1.25	0	.87	.72	.57	.56
7	.72	1	1.03	1.00	.74	.75
8	1.00	1	1.20	1.21	.89	.93
9	1.21	0	.99	.93	.74	.74
10	.93	1	1.12	1.10	.87	.89
11	1.10	0	.95	.89	.73	.72
12	.89	1	1.05	1.02	.84	.84
13	1.02	0	.91	.85	.72	.70
14	.85	1	.99	.96	.82	.80
15	.96	1	1.07	1.05	.90	.90
16	1.05	0	.96	.92	.80	.78
17	.92	1	1.03	1.00	.88	.87
18	1.00	0	.93	.89	.79	.76
19	.89	1	.99	.96	.86	.84
20	.96	1	1.05	1.03	.92	.92

The last four columns of Table 1 contain four different estimates of latent trait status that were calculated after each item was administered. The fourth column contains (approximate) maximum-likelihood estimates of θ. A maximum-likelihood estimate of θ corresponds to the latent trait location at which the observed item responses have maximum probability of occurrence. The probability of a set of item responses, given some fixed value of θ and the item parameters, is obtained using the *likelihood function* given in Equation 3.

$$L_v(\theta) = \prod_g [P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}] \qquad [3]$$

This equation assumes that the responses of a given person to different test items are independent of one another. The operator Π indicates that a serial product is to be taken over the test items administered up to that point ($g=1,2,\ldots,k$).

After each item was administered, Equation 3 was evaluated at 101 equally spaced θ values in the interval from θ=-5.00 to θ=+5.00 and the largest of the 101 likelihood values was identified. Then, a quadratic function was fitted to this largest likelihood value and the two likelihoods adjacent to it. The value of θ corresponding to the maximum of the quadratic function was used as the "MAXL." estimate. Under most conditions, the estimate of θ obtained in this manner is a good approximation to the estimate that would be obtained if we used methods of numerical analysis to search for a root of the log-likelihood function's first derivative. The procedure described here has the advantage that it is not subject to local maxima in the interval between θ=-5.00 and θ=+5.00.

As an aside, we may note that the interval between -5.00 and +5.00 will contain at least 96% of the θ estimates in any group that is used to parameterize test items. This is because latent trait item parameterization procedures scale the theta metric such that the mean θ estimate equals zero and the standard deviation among the estimates is 1.0 (again, the Rasch model provides an exception to this general result), and by virtue of Tchebycheff's inequality which states that the proportion of cases which fall more than S standard deviations from the mean cannot exceed $(1/S^2)$ in any distribution (Hays, 1973, p. 253). If the distribution is peaked and unimodal, virtually all of the θ estimates will be between -5.00 and +5.00.

Figure 7
Relative Likelihood and Posterior Probability Curves After 1 Item

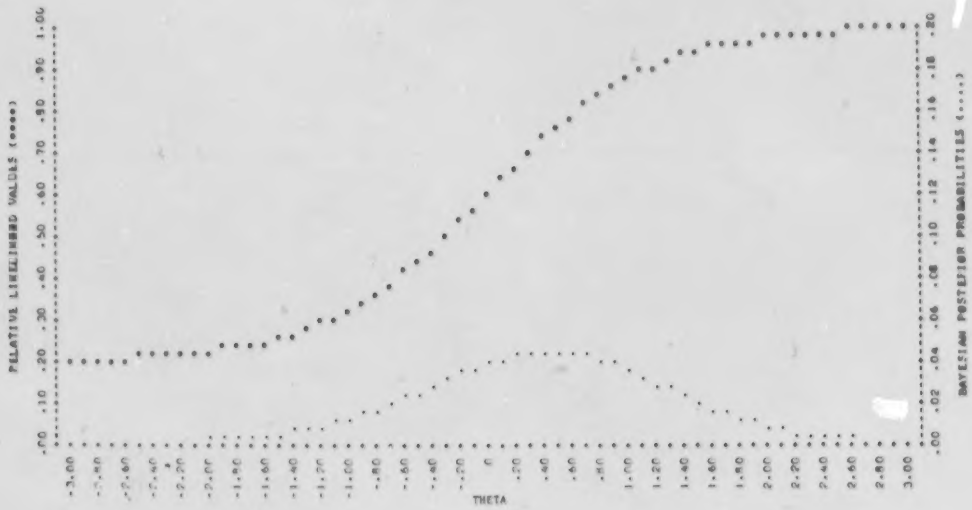


Figure 8
Relative Likelihood and Posterior Probability Curves After 2 Items

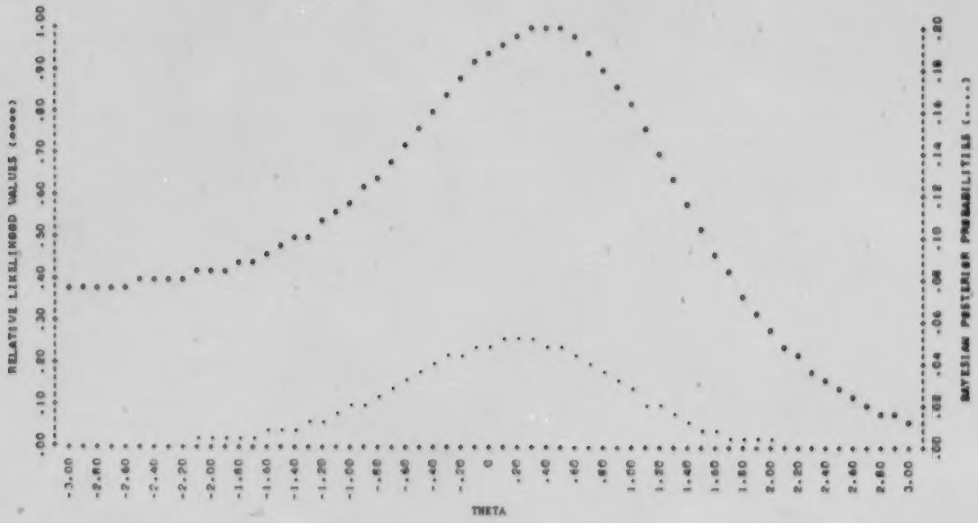
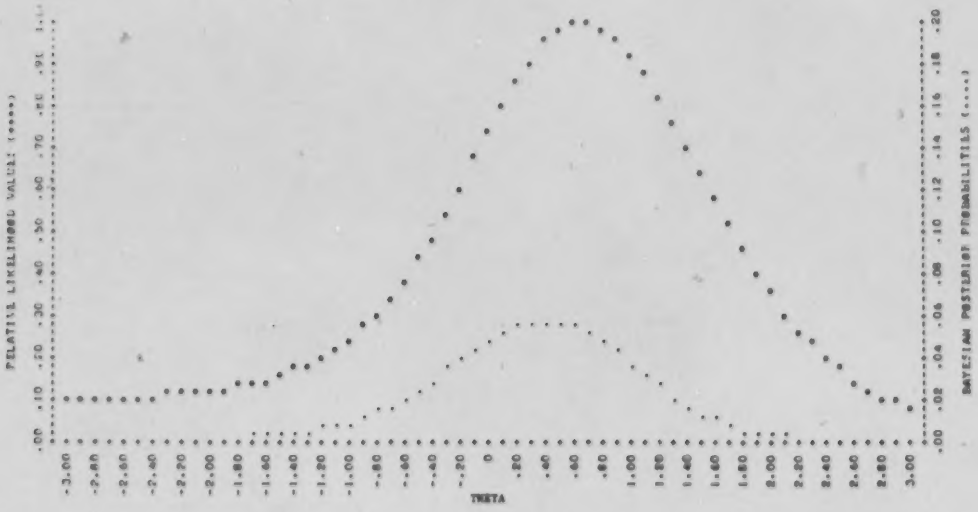


Figure 9
Relative Likelihood and Posterior Probability Curves After 3 Items



Figures 7, 8, and 9 show graphs of the data likelihood function in the interval from θ=-3.00 to θ=+3.00 following the administration of 1, 2, and 3 items, respectively. For plotting purposes, the raw likelihood values were expressed relative to the largest likelihood value in the interval θ=-5.00 to θ=+5.00 and then rounded to the nearest .02. As can be seen in Equation 3, after one item is administered the likelihood function corresponds to either $P_1(\theta)$ or $Q_1(\theta)$, depending on whether a keyed or non-keyed response is emitted. (Compare Figure 7 and Figure 3.) The MAXL. estimate after getting the first item "correct" is +5.49. Actually, since $P_g(\theta)$ is strictly increasing in θ, the estimate should be θ=+∞, but a finite estimate is certainly more reasonable. After "missing" the second item, with b_2=1.00, the peak of the likelihood curve occurs near θ=+.36 (Figure 8). After the third item, the peak occurs near θ=.67 (Figure 9).

A "weighted-by-likelihoods" (W.B.L.) estimate of latent trait status appears in the fifth column of Table 1. The W.B.L. estimates were obtained by taking a weighted average of 101 equally spaced θ values in the interval from θ=-5.00 to θ=+5.00. The weights used were the data likelihoods at each θ value. That is,

$$\text{W.B.L. EST.} = [\sum_{\theta}(L_v(\theta)\ \theta)]/[\sum_{\theta}(L_v(\theta))] \qquad [4]$$

where θ takes on the values -5.00, -4.90, ..., +5.00. The W.B.L. estimate is influenced by the entire set of 101 likelihood values instead of just the three largest values.

The MAXL. and W.B.L. estimates can differ considerably when only a few items have been administered, as can be seen in Table 1. Inspection of the relative likelihood curve in Figure 8 shows why these two estimators differ after two items have been administered. The W.B.L. estimate is lower due to the fact that the left tail of the likelihood curve is high relative to the right tail. Table 1 also shows that the MAXL. and W.B.L. estimators become more similar as the number of items administered increases. To my knowledge, the W.B.L. estimator has not been proposed previously. I plan to study its characteristics in future research.

At this point I can outline the procedure by which item b_g values were determined during the simulated test. The general rule followed was: Let the next item have a difficulty level equal to the current value of the W.B.L. estimator, except that in no case shall the new b_g value be more than 1.00 units from the immediately preceding b_g value. Thus, as can be seen in Table 1, item difficulties changed by 1.00 until the third item had been administered and the W.B.L. estimate was .18. After this, each item difficulty corresponded to the value of the W.B.L. estimate following the preceding item. In actual practice we seldom find an item with b_g exactly equal to the

current estimate of latent trait status. In such cases we select an item that has b_g close to the desired value.

Columns six and seven of Table 1 contain two Bayesian estimates of latent trait status. Given a specified form for the continuous distribution of latent trait scores in a population (i.e., the *prior probability density function* of theta), the item parameters for the items administered, and a vector of item responses (u_g values), one can, in principle, derive the *posterior probability density function* of theta using the inverse probability rule of Bayes (Hays, 1973, p. 819). In practice, it becomes difficult to obtain analytic expressions for the posterior theta distribution unless the prior distribution and the data likelihood function take on certain restricted forms. To avoid such difficulties, I have proposed the following approximate procedure.

First, we approximate the continuous prior density function of theta with a discrete probability distribution in which the probabilities are concentrated at 101 equally spaced points along the theta continuum. Thus, for example, the area under the prior density curve between θ=-.05 and θ=+.05 is assigned to the point θ=.00. This is done for θ=-5.00, -4.90, ..., +5.00. Areas beyond θ=-5.05 and θ=+5.05 are assigned to the points θ=-5.00 and θ=+5.00 respectively. (These extreme tail areas should be trivially small. If they are not, one can either shift or extend the region of the theta continuum in which the procedure is applied.) Next, data likelihoods are generated at the same 101 values of θ using Equation 3. The prior probabilities, $f(\theta)$, and the data likelihoods, $L_v(\theta)$, are then entered into Equation 5 in order to determine the posterior probability of each given θ value.

$$P(\theta|v) = [L_v(\theta)\ f(\theta)]/\sum_{\theta}[L_v(\theta)\ f(\theta)] \qquad [5]$$

The resulting 101 posterior probabilities provide a discrete approximation to the continuous posterior distribution of theta. Finally, the mean of the discrete posterior distribution is obtained with Equation 6 and this value is referred to as the "SBAYES" (simplified Bayesian) estimate at that stage of the testing procedure.

$$\text{SBAYES EST.} = \sum_{\theta}[P(\theta|v)\ \theta] \qquad [6]$$

Figures 7, 8, and 9 show three of the posterior probability distributions that were generated with the SBAYES procedure when the prior distribution of latent trait scores was specified to be a normal density function with zero mean and unit variance. The first three SBAYES estimates in Table 1 are the means of these discrete distributions.

The "OBAYES" (Owen Bayesian) latent trait estimates that appear in column seven of Table 1 were obtained using a procedure described

by Owen (1975). While Owen described a method for estimating latent trait status and a method for selecting test items, I have utilized only his estimation procedure here. Owen introduced his procedure in the context of a *three-parameter normal ogive* latent trait model. The close similarity of this model to the logistic model given in Equation 1 allows its application here.

The OBAYES procedure has two drawbacks. First, it is limited to prior distributions that follow a normal density function. The SBAYES procedure described above can accept any type of prior distribution. Second, the OBAYES procedure is order dependent. That is, if a set of items is administered and the item responses are recorded, then the value of the OBAYES estimator will depend partly on the order in which the items are processed by the scoring procedure. The OBAYES procedure (implicitly) generates an updated prior distribution after each item is scored and then combines this new prior distribution with the likelihood function for the response to the next item. This in itself would not make the OBAYES procedure order dependent but, in order to simplify the mathematics, Owen proceeded as if each updated prior distribution could be described by a normal density function. This approximation introduces a small amount of inaccuracy into the estimation process and makes the procedure order dependent. The SBAYES procedure utilizes the same prior distribution at each stage of the testing process and updates the data likelihood function after each item is scored. The SBAYES procedure is not order dependent.

After administering a single item, SBAYES and OBAYES estimates generally agree to three decimal places when the prior distribution is a normal density function. Since the OBAYES estimate is optimal in this particular situation, we can view this level of agreement as an indication that very little inaccuracy is introduced by the discrete approximations in the SBAYES procedure. When more than one item has been administered, or when the prior distribution specified for the SBAYES procedure is non-normal, the two estimation methods will not necessarily agree, as can be seen in Table 1.

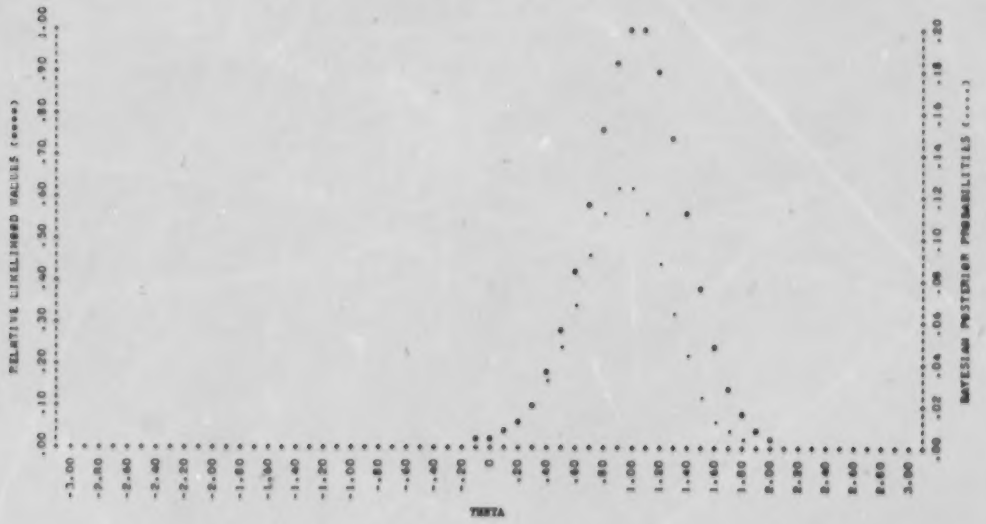
Figure 10 shows the relative likelihood and posterior probability curves that resulted after 20 items had been administered. The likelihood curve peaks near θ=1.05 and the posterior probability distribution has a mean of .92 (see Table 1). Both the likelihood curve and the posterior probability curve have shifted to the region of the theta continuum near θ=1.00, and both curves have become more peaked. In fact, as test length (k) approaches infinity, both of these curves approach a vertical line (i.e., a single-valued distribution) located at the value of θ that is generating the item responses.

Note in Table 1 that the Bayesian estimates of θ tend to stay closer to θ=.00 than the other estimates throughout the testing process. This is because Bayesian estimators are "drawn toward" the high density region of the prior distribution. This is appropriate when one's objective is to minimize squared errors of estimation in the population specified by the prior distribution. Unfortunately, for tests of moderate length, we must accept a certain amount of

bias at the tails of the theta distribution in order to achieve this minimization (McBride & Weiss, 1976).

For moderate k, the maximum-likelihood estimator can also be biased. However, for a given value of k and values of θ deviant from the high density region of a peaked prior distribution, the maximum-likelihood estimator will tend to be less biased than the Bayesian estimator. The Bayesian estimator's bias can be reduced by increasing k as the estimate of θ deviates from the high density region of the prior distribution. This can be done readily in an adaptive testing situation.

Figure 10
Relative Likelihood and Posterior Probability Curves After 20 Items



An interesting relationship exists between the MAXL. and W.B.L. estimators and Bayesian estimators. If one applied the SBAYES estimation procedure and specified that the prior distribution of theta was rectangular in the interval θ=-5.05 to θ=+5.05, then the SBAYES estimate of θ, as determined by Equation 6, would be identical to the W.B.L. estimator. Moreover, the MAXL. estimate would closely approximate the mode of the Bayesian posterior probability distribution. Thus, we see that all four types of latent trait estimator that have been presented here can be viewed as Bayesian estimators. The MAXL. estimator is a Bayesian modal estimate of θ when the (implicit) prior is restricted to a rectangular form, the W.B.L. estimator is a least-squares estimate of θ when the (implicit) prior is restricted to a rectangular form, and the OBAYES estimator is a least-squares estimate of θ when the (explicit) prior is restricted to a normal form. The SBAYES procedure is the only one of the four

methods that does not restrict the form of the prior distribution. By virtue of this flexibility, I am inclined to recommend the SBAYES estimation procedure as the most widely applicable of the four methods.

Total test information. Birnbaum (1968, p. 454) has defined the *information function of a test* as

$$I(\theta) = \sum_{g} [I(\theta, u_g)]. \qquad [7]$$

This function is the sum of the constituent item information functions and defines the maximum amount of information that can be extracted from a set of items. The amount of information actually extracted depends on how the items are scored.

Figure 11 shows a graph of the test information function for the 20 items administered in our simulated adaptive test. It was obtained by evaluating Equation 7 at equally spaced points along the theta continuum in the interval from θ=-3.00 to θ=+3.00. This curve shows the maximum amount of information available from the set of items, regardless of how they are scored. The curve peaks near θ=1.00, thus indicating that this set of items provides maximum discrimination between individuals whose latent trait locations fall near θ=1.00. The maximum value of the curve is about 9.00.

Figure 11
Information Curve for 20-Item Adaptive Test

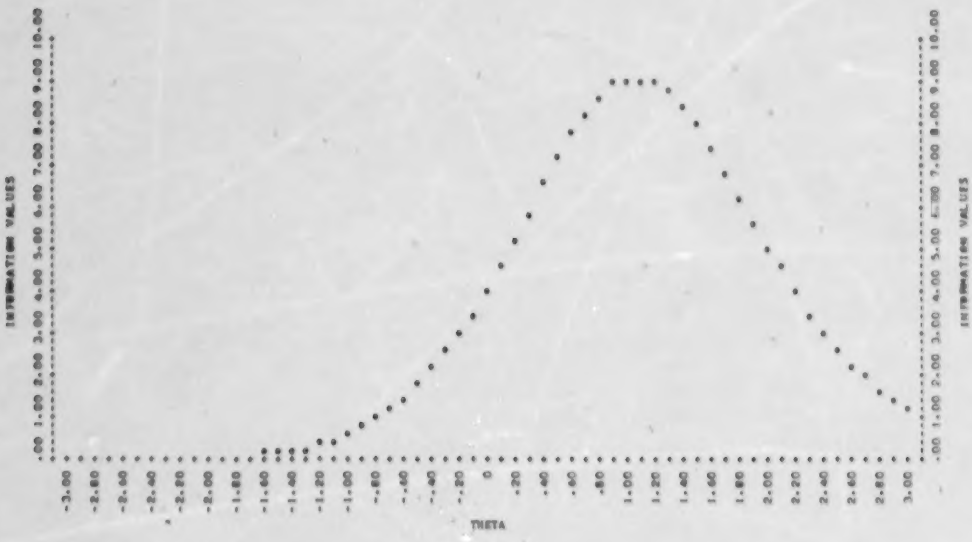


Figure 12
Information Curve for 20-Item Rectangular Test ($-3.0 \leq b \leq +3.0$)

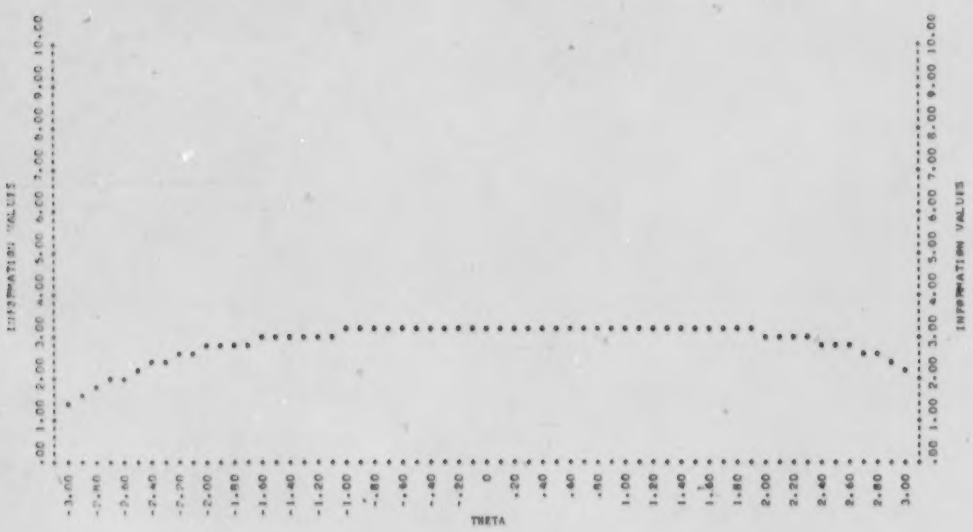


Figure 13
Information Curve for 20-Item Peaked Test (b=0.0 for all items)

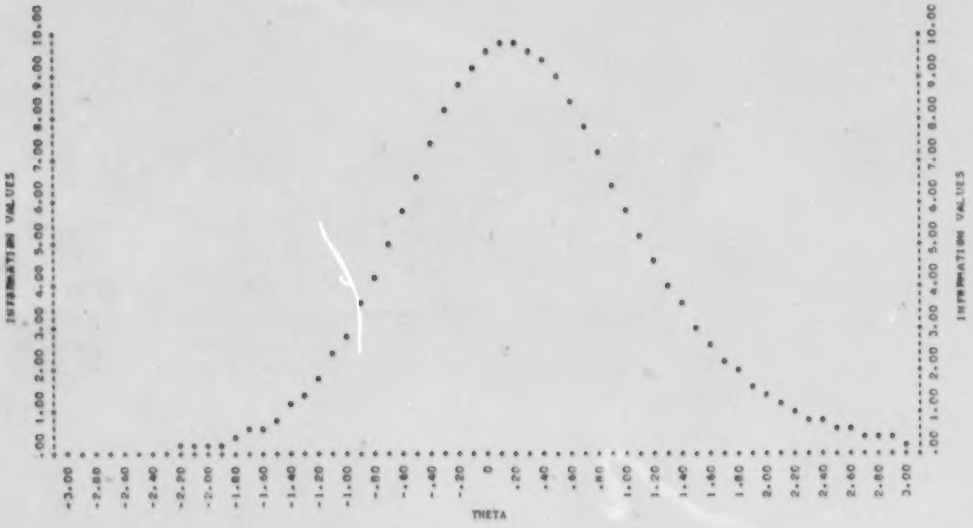


Figure 12 shows a graph of the test information function for a set of 20 items having a_g=1.0, c_g=.20, and b_g values equally spaced in the interval from -3.00 to +3.00 (i.e., b_g=-3.00, -2.68, -2.37, ..., +3.00). This would commonly be referred to as a "rectangular" distribution of item difficulties. This test provides a fairly uniform level of information across a broad range of the theta continuum. Unfortunately, the level of information is relatively low. The curve attains its maximum value of about 3.20 in the interval $-1.00 \leq \theta \leq 1.90$.

Figure 13 shows a graph of the test information function for a set of 20 items having a_g=1.0, c_g=.20, and b_g=0.0 for all items. This is a "perfectly peaked" test. The shape of this information curve is rather similar to the curve in Figure 11, but it is shifted to the left. The curve in Figure 13 attains its maximum value of 9.80 near θ=.16. At θ=1.00, the value of this information curve is about 5.80.

Figures 12 and 13 represent two rather idealized non-adaptive tests. We have seen that both of these tests deliver less information at θ=1.00 than the items selected by the adaptive testing procedure. What is the implication of this result? If we had to order our imaginary testee relative to other individuals falling at nearby θ values, we would make fewer mistakes if we used θ estimates derived from the adaptive test's items than if we used estimates derived from either conventional test's items.

Concluding Remarks

I have presented several procedures for estimating latent trait status. I have also suggested that adaptive testing procedures often can provide more accurate estimates of latent trait status than conventional tests. Though there is no necessary connection between latent trait theory and adaptive testing, there is a strong natural impetus toward their joint application. Latent trait theory provides adaptive testing with a coherent theoretical foundation. It is a guide to procedures for designing and scoring adaptive tests. On the other hand, adaptive testing offers us the opportunity to take maximum advantage of the potentialities of latent trait theory. Adaptive testing offers the possibility that test reliability and/or validity might be increased or that some specified level of reliability or validity might be achieved in less testing time. We are now at a fortunate point in time where both a new type of test theory and a new type of testing technology are available. Their joint effect might possibly exceed the sum of the two parts.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968, Chapters 17-20.

Bock, R.D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37(1), 29-51.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40(1), 5-32.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10(3), 159-170.

Hays, W.L. Statistics for the social sciences. New York: Holt, Rinehart, & Winston, 1973.

Lord, F.M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

McBride, J.R., & Weiss, D.J. Some properties of a Bayesian adaptive ability testing strategy. (Research Report 76-1). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1976.

Nunnally, J.C. Psychometric theory. New York: McGraw-Hill, 1967.

Owen, R.J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70(350), 351-356.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Neilson & Lydiche (for Danmarks Paedagogiske Institut), 1960.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39(1), 111-121.

Wright, B.D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Adaptive Testing and the Problem of Classification

by
C. David Vale
University of Minnesota

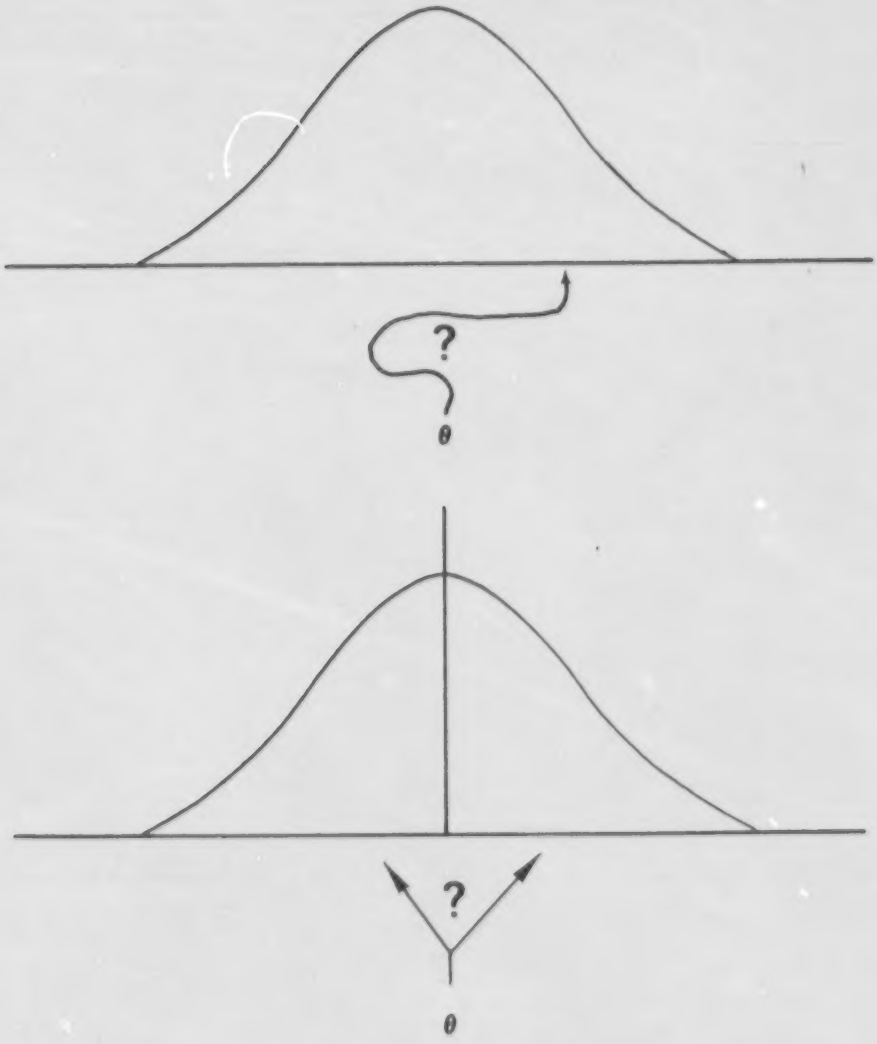
Two basic goals in the use of ability tests are measurement and categorization. When we use a test for measurement, we want to accurately determine where a testee's ability falls on the latent ability continuum. When we use a test for categorization, we want to determine on which side of a cutting score (or between which cutting scores) a testee's ability lies with as little chance of making a mistake as possible. Beyond that coarse determinization, we don't care where his ability is. These two approaches are illustrated in Figure 1.

In this paper I will discuss the categorization of abilities. Specifically, I will explicate the general goals of categorization; discuss how, in theory, these goals can be optimally met; and present and compare some alternative means that may practically be used to achieve these goals.

I mentioned that the goal of categorization is to determine on which side of a cutting score, or between which of several cutting scores, a testee's ability falls with a minimal probability of being in error. There are really two probabilities that we need to discuss; one is the probability of being in error for a single testee (i.e., at a specific ability level) and the other is the expected probability of being in error across a group of testees. The first probability is a function of the test, the testee's ability level and the placement of the cutting score (we will discuss only the case of one cutting score for the time being). For a given test of fixed length, the probability of making an error of classification is usually high if his ability level (Θ) is near a cutting score and lower if his ability is distant from the cutting score. This conditional probability of misclassification [$P(M|\Theta)$] is described by a function like that shown in Figure 2. The second probability, the expected probability of misclassification for a group of testees [$P(M)$], is a function of the conditional probability function and the distribution of ability within the group. For a large group with abilities distributed $N(0,1)$, the expected probability is given by Equation 1.

This research is supported by Contract N00014-76-C-0243, NR 150-382, with the Personnel and Training Research Programs, Office of Naval Research, David J. Weiss, Principal Investigator.

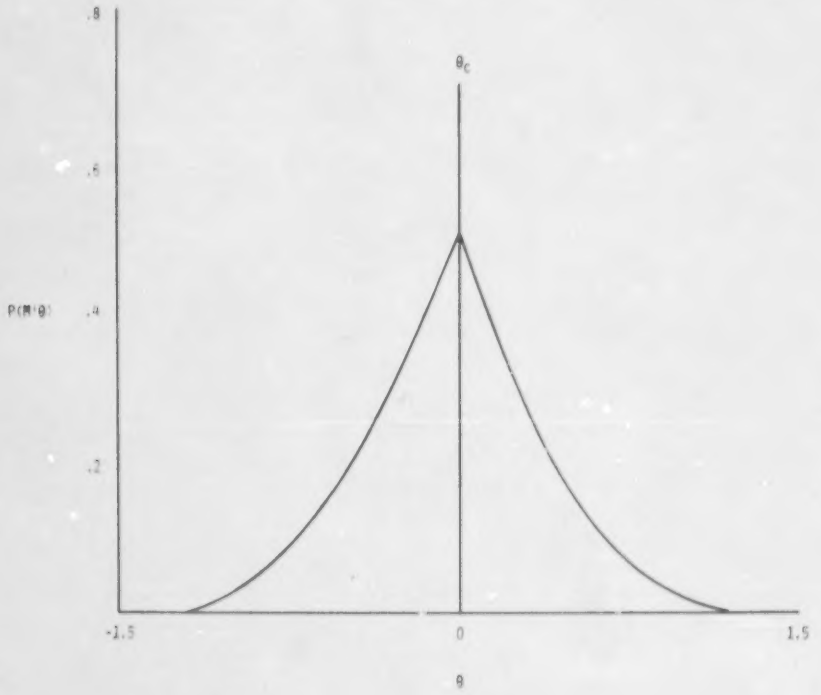
Figure 1
Measurement Versus Classification



$$P(M) = \int_{-\infty}^{\infty} P(M|\Theta)\ \phi\ (\Theta)\ d\Theta$$

where $\phi\ (\Theta) = e^{\frac{-\Theta^2}{2}} / \sqrt{2\pi}$ [1]

Figure 2
A Conditional Probability of Misclassification Curve



In practical situations, it is usually this expected probability that we want to minimize. But it is very difficult to do this explicitly and approximation strategies are used. In this paper I will compare results of monte carlo simulations of conventional testing strategies with those of an adaptive measurement strategy. A brief description of the simulation methods used is presented in the Appendix.

The model of making classification decisions I will assume is as follows:

1. Items are selected from a larger pool of items and administered to a testee.

2. From responses to these items, an estimate of a testee's ability is obtained. (The scoring method used in this study was Owen's (1975) Bayesian procedure starting with an $N(0,1)$ prior.)

3. Testees are classified into that category in which their ability estimates fall.

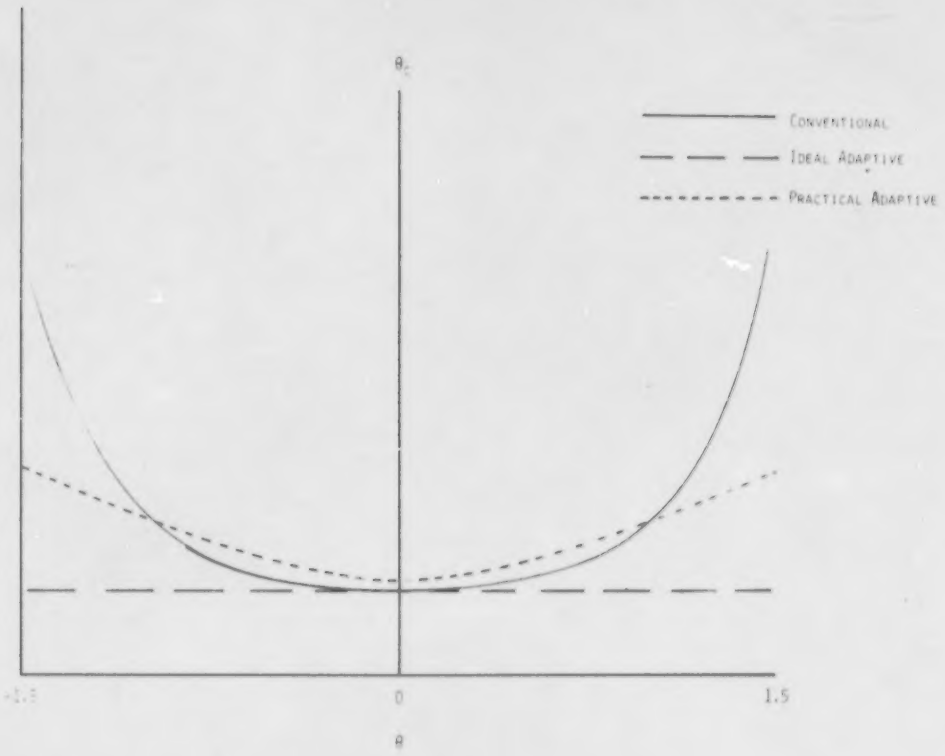
To simplify the analyses and interpretations it was assumed that, for item selection, an infinitely large item pool, containing items of all difficulties with their discriminating powers fixed at a constant level, was available from which to choose. It was further assumed that it was not possible to answer these items correctly by guessing. These assumptions reduced the problem of item selection to determining how difficult the item should be. Finally, to make determination of the expected probability of misclassification possible, ability was assumed distributed $N(0,1)$.

The first situation I will discuss is conceptually the simplest categorization situation to investigate. In this situation, there is one cutting score placed in the middle of the ability distribution at $\theta_c=0.0$. The best conventional test for making this decision is one with all of its items peaked at $b=0.0$. Due to a misinterpretation of the concept of information discussed earlier by McBride and Sympson, it is sometimes thought that this is the best test, adaptive or conventional, that can be constructed for this situation. The logic of the argument goes something like this: "Information is an index of a test's capacity to discriminate between different abilities. Since, in this case, we only need to discriminate those ability levels above the cut from those below the cut, we only need information at the level of the cut. Information at this level can be maximized by peaking the items at $b=0.0$."

A more rigorous interpretation of information is as the reciprocal of the squared standard error of measurement produced by an unbiased ability estimate at a given level of ability. Figure 3 shows curves representing standard error of measurement functions for three ten-item tests: a peaked conventional test with all items having difficulty $b=0.0$, an ideal adaptive test with all items having $b=\theta$, and a practical adaptive test with items having difficulties at the estimated ability level at each stage. The conventional test provides a low error level at $\theta=0.0$ but higher error levels distant from that point. The ideal adaptive test provides this low level of error at all ability levels but is unrealistic because in order to implement it, we need to know a testee's ability level before we test him. A practical adaptive test provides a standard error function lower than that of the con-

ventional test at ability levels distant from Θ=0.0 but relatively higher near Θ=0.0.

Figure 3
Standard Error of Measurement Curves for Three Tests



Assuming errors of measurement at a level of Θ are distributed $N(\Theta, SEM^2)$, the probability of misclassifying an individual is given by Equations 2 and 3.

$$P(M|\Theta) = 1 - \Phi\left[\frac{|\Theta_c - \Theta|}{SEM}\right] \quad [2]$$

$$= 1 - \Phi[\sqrt{I(\Theta)\ (\Theta_c - \Theta)^2}] \quad [3]$$

where Θ_c is the cutting score,

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt,$$

and $I(\Theta)$ is the test information function evaluated at Θ.

It can be shown from Equation 2 that when Θ_c is fixed, $P(M|\Theta)$ is a monotonic increasing function of the standard error of measurement. Thus, the ordering of the three types of tests on $P(M|\Theta)$ is the same as their ordering on conditional standard errors of measurement at any level of Θ. (I present error of measurement functions here simply because in the theoretical case where we assume the existence of an unbiased score, the $P(M|\Theta)$ curves are difficult to distinguish in a figure.) It can be seen from these curves that a practical adaptive test can provide a lower expected probability of misclassification if its approximation to the ideal adaptive test is close enough. Whether an adaptive test can approximate the ideal well enough is, of course, an empirical question.

Figure 4
Conditional Probability of Misclassification, a=1.0

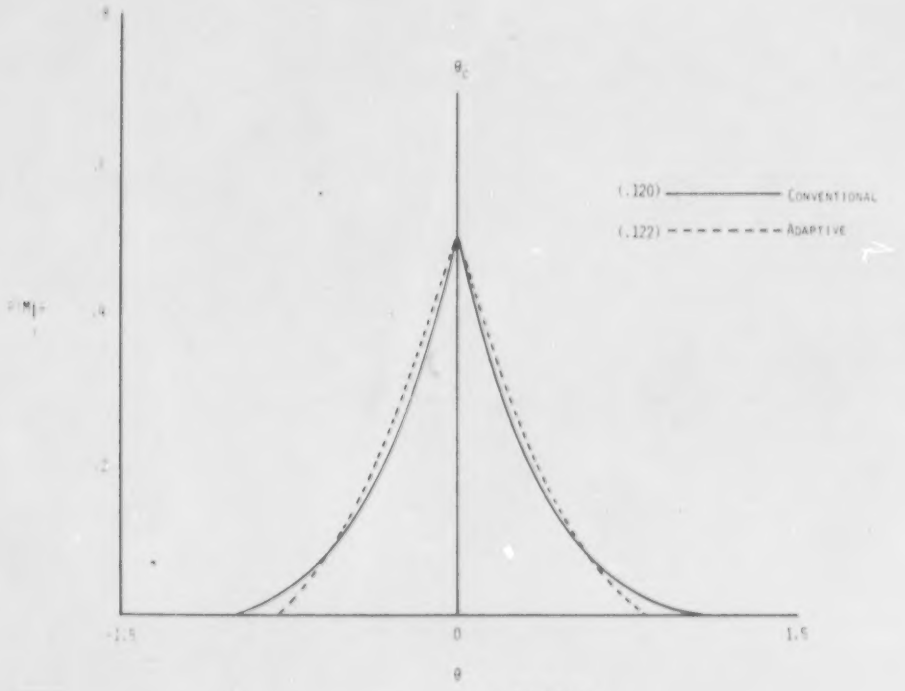


Figure 4 presents the $P(M|\Theta)$ curves for a ten-item conventional test, with difficulties peaked at b=0.0, and a ten-item Bayesian adaptive test (Owen, 1975), both with item discriminations fixed at a=1.0. Both tests were scored by Owen's Bayesian scoring method. The curves appear very similar, being high near the cut (indicating a high probability of making an error) and low distant from the cut. The conventional test allows somewhat better decisions near the cut while the adaptive test allows relatively better decisions distant from the cut. These trivial differences in the conditional probability of misclassification function yield a trivial difference between the expected probability of misclassification values for the two strategies (.120 for the conventional test and .122 for the Bayesian test).

Figure 5
Conditional Probability of Misclassification, a=2.0

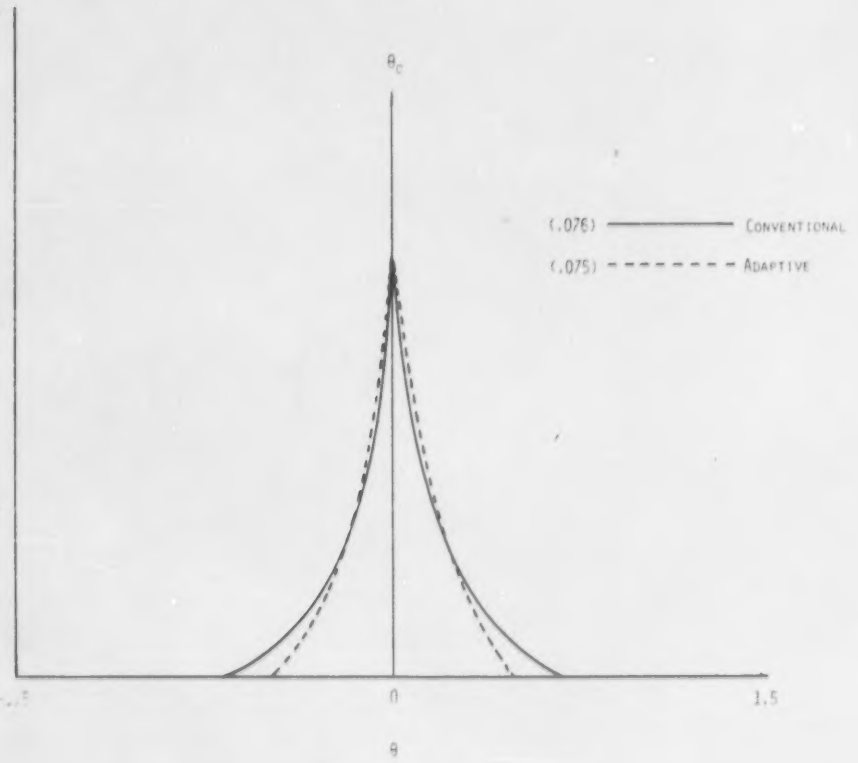


Figure 5 shows $P(M|\Theta)$ curves for the same strategies with item discriminations of a=2.0. The same general comments on shape and relative level apply, except the differences distant from the cut are more pronounced. Due to the shape of the ability distribution (i.e., $N(0,1)$), however, small differences near the cut are as important in the determination of the expected probability of misclassification as large differences distant from the cut, and the difference in expected probability is still trivial (.076 versus .075).

Figure 6
Conditional Probability of Misclassification, a=3.0

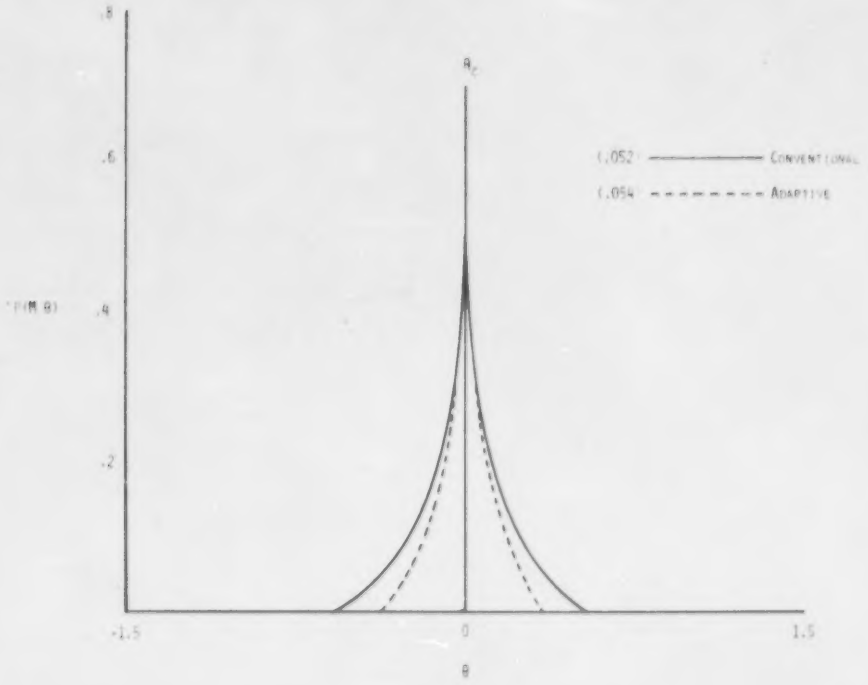


Figure 6 shows curves for tests with high item discriminations (a=3.0). Again the same general comments regarding shape and level of the curve apply and the difference in expected probability of misclassification is equally trivial (.052 versus .054).

These results suggest that an adaptive test makes classification decisions about as well as a conventional test in this simple case where a conventional should fare better in comparison to an adaptive test. When there are more cutting scores, an adaptive test should provide relatively better decisions than was demonstrated in this situation.

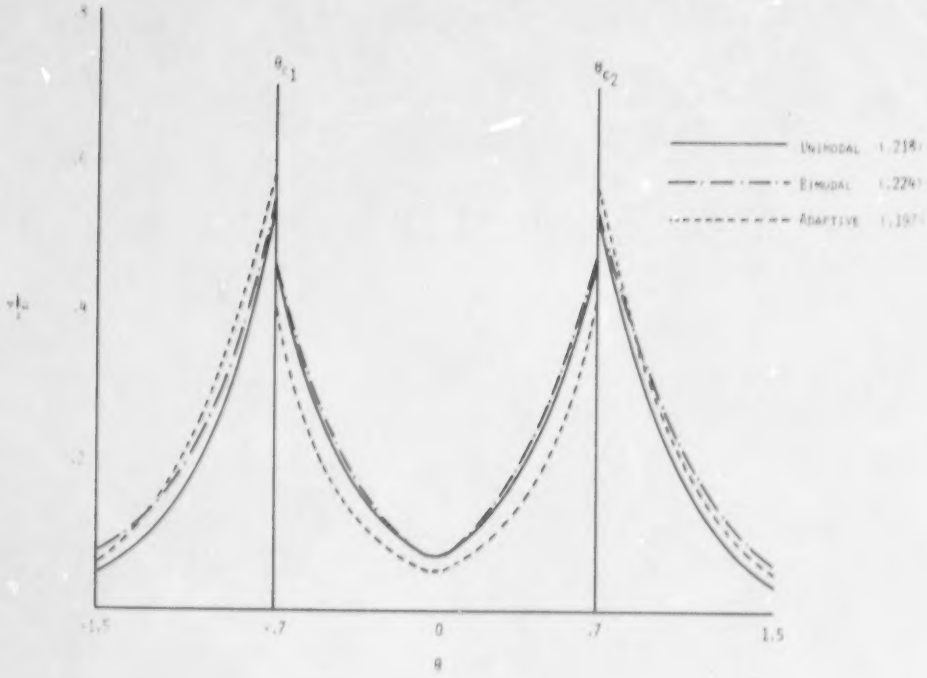
Design of conventional tests is more complicated, however, when the cutting scores deviate from the center of the ability distribution. A given increase in information, which corresponds to a given decrease in standard error, has its greatest effect on the conditional probability of misclassification at ability levels near a cutting score. This suggests that we peak items at the cutting scores. But a given reduction in conditional probability of misclassification has its greatest effect on the expected probability of misclassification at levels of ability where most of the people are located. This, assuming $\Theta \sim N(0,1)$, suggests peaking the item difficulties at b=0.0. So when the cutting score is at some Θ not equal to 0.0, the two suggestions are in conflict. The optimal point(s) to peak the difficulties will be some function of the location of the cutting scores, the discriminating powers of the items, and the underlying ability distribution. Determination of such an optimal design of a conventional test is beyond the scope of this paper and instead of comparisons with this test, comparisons of some standard conventional test designs with an adaptive test will be presented.

Assume that we have two cutting scores, one at Θ_c=-.7 and the other at Θ_c=.7, and that all errors of misclassification are equivalent in terms of importance. One classical approach to designing a conventional test involves peaking half of the items at each of the two cutting scores, where the fine distinctions need to be made. We will refer to such a test as a bimodal conventional test. Another approach is to peak all items at b=0.0. We will refer to this test as a unimodal conventional test.

Figures 7 through 9 present the conditional probabilities of misclassification for each of the unimodal and bimodal conventional and Bayesian adaptive testing strategies at three levels of item discrimination. Figure 7 shows the curves for the case when a=1.0. There is little suggestion of which strategy is better but an interesting discontinuity is observed for estimates from all scoring strategies at the cut points. This characteristic is due to the fact that, for finite length tests (which includes 10-item tests like those used here), the Owen's Bayesian score is biased (i.e., the expected value of the score at a given level of Θ is not Θ). Specifically, in this case, the Bayesian score is biased in the vicinity of the cutting scores toward the center of population ability distribution at Θ=0.0. This causes more testees

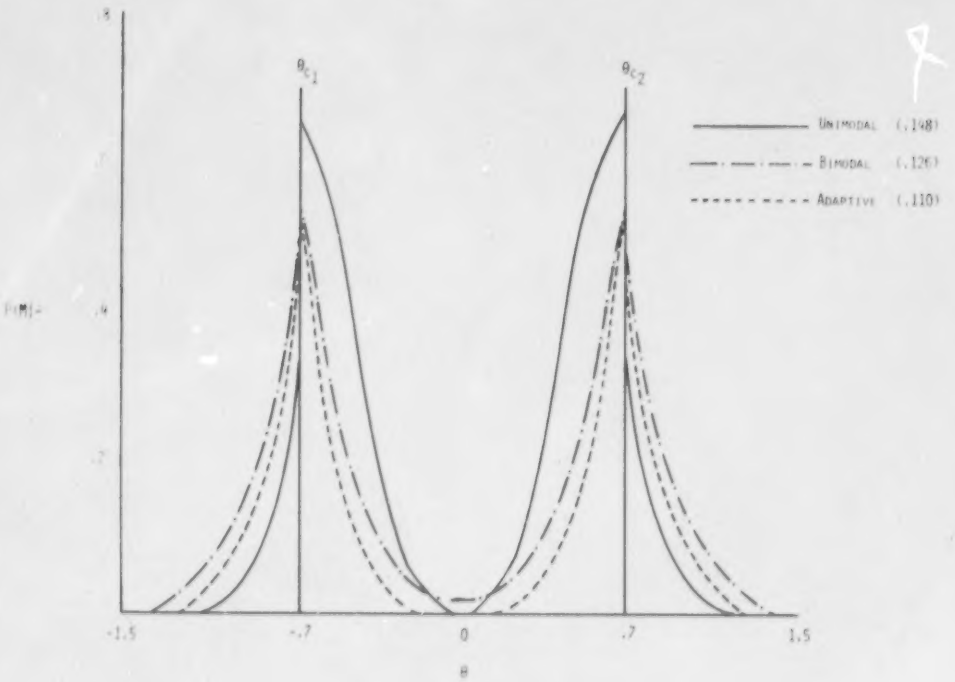
to be classified into the middle interval than would be by an unbiased score. The effect is that fewer errors of classificat... are made for ability levels in the middle interval and more are made for individuals in the two extreme intervals. Comparing expected probabilities of misclassification, the adaptive test yields the lowest probability (.197) and the bimodal conventional, the highest (.224).

Figure 7
Conditional Probability of Misclassification, a=1.0



It is difficult to say in this case, however, whether the adaptive test provides a lower expected probability of misclassification because it makes better decisions or because it is conservative in terms of classifying into the extreme categories and thus makes fewer errors at central ability levels where more individuals' ability levels lie.

Figure 8
Conditional Probability of Misclassification, a=2.0

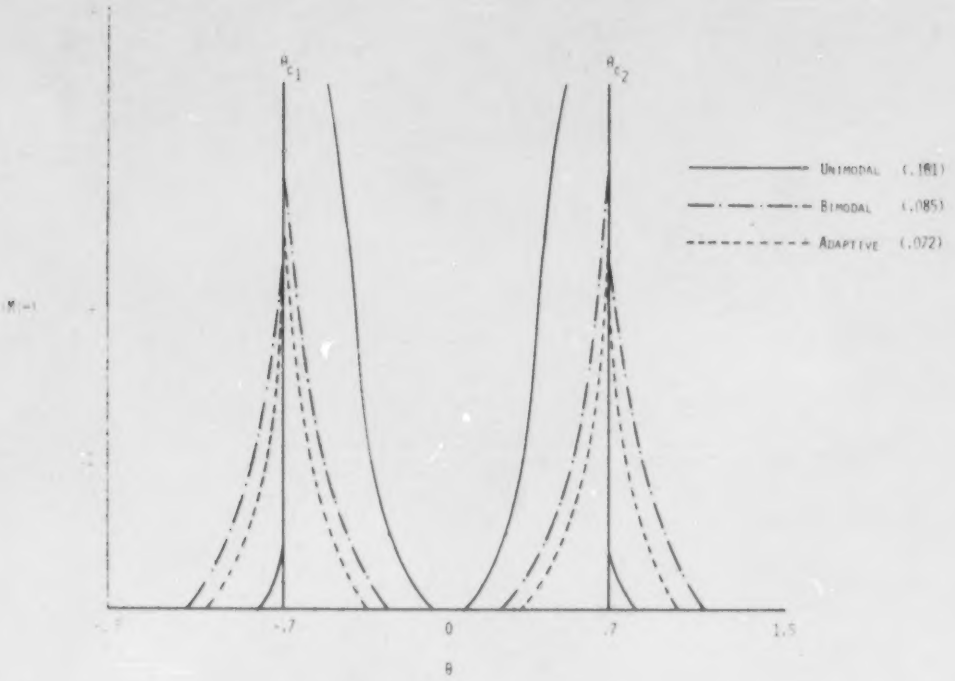


When a=2.0 (Figure 8), the unimodal conventional test shows pronounced discontinuity, suggesting that scores are too extreme near the cuts. The adaptive test provides the smallest conditional probabilities of misclassification over most of the ability range. It makes a few more errors in the extreme intervals than does the unimodal conventional test, but the unimodal test's superiority is offset by extreme error rates in the middle interval. In terms of expected probabilities of misclassification, the adaptive test is again superior [$P(M)$=.110]. With an expected probability of misclassification of .126, the bimodal conventional test, its nearest competitor, is expected to make 1.15 times as many errors of classification.

Since the adaptive test is superior to the bimodal conventional test over the entire range of ability in terms of conditional probabilities of misclassification, its superiority in expected probability of misclassification is expected. And, considering

the low level of bias evident in the figure, the actual values are probably nearly correct.

Figure 9
Conditional Probability of Misclassification, a=3.0

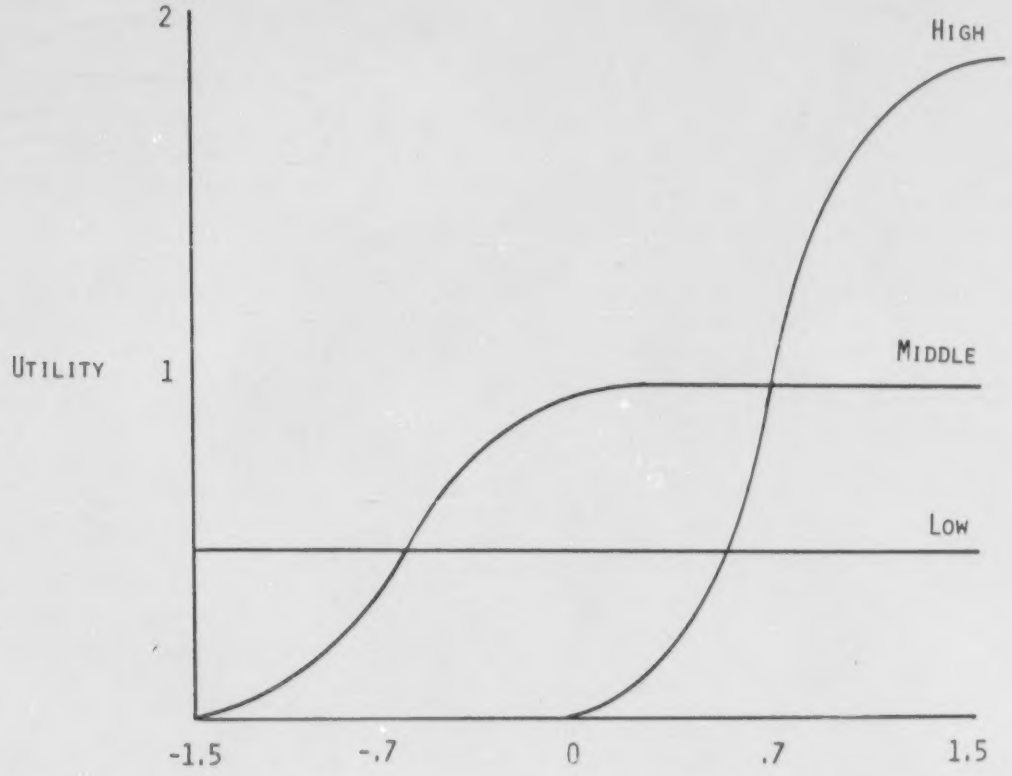


When a=3.0, as shown in Figure 9, the same general comments regarding comparisons between the conditional probability of misclassification curves apply. The expected probability of misclassification for the bimodal conventional test (.085) is 1.18 times as large as that of the Bayesian adaptive test (.072). As was mentioned earlier, however, items this discriminating are rare in practice.

It is tempting to take these values at this point and say that adaptive testing can greatly reduce errors of classification (assuming a 13% reduction is great) in a practical classification situation. But note that we are discussing errors of classification on a latent ability continuum rather than an observable success-

failure continuum. This fact is important because two individuals, one with ability slightly above a cut on the latent continuum and the other with ability slightly below the cut, probably have a trivial difference between their probabilities of success on a job. If both are classified above the cutting score, however, one will be considered a "hit" and the other a "miss" in the current

Figure 10
Conditional Utilities for each of Three Decisions



scheme. In order to assess the value of an adaptive strategy, utility functions of Θ for each decision must be specified. As an example of such utility functions, consider the following:

For three classifications, low, middle, and high, the three utility functions are:

$u_{low} = .5$

$$U_{medium} = \Phi(3.0(\Theta+0.7))$$

$$U_{high} = 2.0(\Phi(3.0(\Theta-0.7))$$

This yields the three utility curves presented in Figure 10 which might give rise to the cutting scores at Θ_i=-.7 and Θ_i=.7 discussed earlier. A practical situation in which these utility functions might arise is as follows: There are three jobs requiring some ability, Θ. One is so easy that almost anyone can do it but when performed satisfactorily, it is of only .5 utility units of value to the company. A second job is fairly easy and 50% of people with Θ above -.7 can perform it satisfactorily. Differences in ability near -.7 make greater changes in the probability of success than do differences around, say, Θ=0.0. Ninety-eight percent of people with Θ above 0.0 will be successful on the job and additional increments in Θ are of trivial importance in predicting job success. Success in this job is worth one unit of value. A third job requires higher Θ to be successful but is worth two units of value when performed satisfactorily.

Choosing the decision yielding the highest expected utility for the estimate of ability, average utilities for the bimodal conventional test and the Bayesian test were, in simulation with items of a=1.0, .808 and .820, respectively. For the same sample of abilities and a=2.0, the utilities were .831 and .849. With a=3.0, the values were .855 and .858. Whether these differences are practically significant depends on what these units of utility mean in a particular context, but such utilities (of which these are only an example) must ultimately be considered in determining the comparative values of conventional versus adaptive testing for classification decisions.

In this paper, I have suggested that adaptive testing may offer significant advantages to an organization involved in making classification (e.g., selection and placement) decisions. Specifically, I have shown that, although a conventional test is as good as an adaptive test when there is one cutting score at the middle of the ability distribution, an adaptive test will provide better categorization when there is more than one. To determine cost effectiveness of adaptive testing in an organization, utility functions will have to be considered, however. Determination of these is beyond both the realm of psychometric theory and the scope of this paper.

APPENDIX

Generation of Probabilities of Misclassification

To generate the conditional probability of misclassification curves presented in this paper, conditional probability of misclassification values was calculated for each of 30 values of Θ equally spaced between Θ=-1.45 and Θ=1.45. Each of these values was calculated by first administering, in simulation (see McBride & Weiss, 1976; or Vale and Weiss, 1975, for a discussion of psychometric simulations), ten-item "tests" to 200 "testees". The means and standard deviations of these estimates were calculated, a normal distribution with these parameters was determined, and the proportion of that distribution falling outside the correct cutting score interval was taken as the probability of misclassification at that level of ability.

To determine the expected probability of misclassification, ten-item "tests" with ability levels randomly sampled from a $N(0,1)$ population of ability levels (the same sample of 2000 ability levels was used for all comparisons). The predicted category for an individual was the score interval in which his ability estimate fell. The true category was the interval in which his true ability fell. An individual was considered misclassified if his estimated category was not the same as his true category. The number of misclassified individuals divided by 2000 was taken as the expected probability of misclassification.

Expected utility was determined by generating 2000 ability estimates the same way as was done for the expected probability of misclassification calculations. The optimal decision to make for an individual was taken as the decision corresponding to the utility function with the highest value at the estimated level of ability. The actual utility was the value of the utility function, corresponding to the decision made, evaluated at the "testees" true level of ability. The expected utility was simply the mean of these 2000 actual utility values.

REFERENCES

McBride, J.R., & Weiss, D.J. Some properties of a Bayesian adaptive ability testing strategy. (Research Report 76-1). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1976.

Owen, R.J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70(350), 351-356.

Vale, C.D., & Weiss, D.J. A simulation study of stradaptive ability testing. (Research Report 75-6). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1975.

Applications of Adaptive Testing in Measuring Achievement and Performance

by
Isaac I. Bejar
University of Minnesota

Introduction

Achievement testing consists of locating individuals on an achievement scale. Usually, to interpret achievement test scores, a transformation is applied to the scores which allows an interpretation in terms of the relative standing of an individual with respect to the norming group. In many instructional settings, this interpretation is not adequate and, as a result, a demand for more concrete kinds of interpretation has emerged. The frequency with which criterion-referenced testing, mastery testing and similar approaches are used is evidence that the suggestion has been welcomed by test users.

What is unique about these testing procedures is that the items that constitute the test are sampled from a population of items which is isomorphic with the objectives of the instructional program on which we want to measure achievement (Shoemaker, 1975). Because of this, it is possible to interpret scores in terms of what the student can do in relation to the objectives of the instructional program.

Undoubtedly, this attention to content is bound to increase the quality of test scores. Today I'd like to describe our efforts at the University of Minnesota to improve achievement testing in general, including criterion-referenced testing approaches, by means of more refined response procedures as well as by adapting the test to the individual.

Background

Most psychometric theory assumes dichotomous scoring; that is, responses are classified as either correct or incorrect. However, knowledge is seldom binary, and by proceeding as if it were, partial knowledge is not given due recognition. If, in fact, partial information is present, then extracting it should lead to more valid and reliable scores.

The research literature, however, does not support the last statement. The results of the typical investigation show that while reliability is usually increased by taking partial knowledge into

---

This research is supported by Contract N00014-76-C-0627, NR 150-389, with the Personnel and Training Research Programs, Office of Naval Research, David J. Weiss, Principal Investigator.

account, the validity of the scores remains the same or even diminishes. Such findings are usually interpreted as evidence against the usefulness of the assessment of partial knowledge. To me, they indicate that something is amiss, for example, that the test and the criterion are not unidimensional.

To illustrate, consider two tests, A and B, measuring a single construct. Both A and B correlate .60 with the construct. This can be summarized as follows:

$$\Lambda = \begin{matrix} \text{Test} \\ \begin{bmatrix} .60 \\ .60 \end{bmatrix} \end{matrix} \begin{matrix} \\ \text{A} \\ \text{B} \end{matrix} \qquad [1]$$

Then the intertest correlation matrix can be expressed as in Equation 2,

$$\Sigma = \Lambda\Lambda' + \Psi, \qquad [2]$$

which in this case becomes Equation 3:

$$\Sigma = \begin{matrix} \Lambda \\ \begin{bmatrix} .60 \\ .60 \end{bmatrix} \end{matrix} \begin{matrix} \Lambda' \\ \begin{bmatrix} .60 & .60 \end{bmatrix} \end{matrix} + \begin{bmatrix} .64 & .00 \\ .00 & .64 \end{bmatrix}$$

$$= \begin{bmatrix} .36 & .36 \\ .36 & .36 \end{bmatrix} + \begin{bmatrix} .64 & .00 \\ .00 & .64 \end{bmatrix}$$

$$= \begin{bmatrix} 1.00 & .36 \\ .36 & 1.00 \end{bmatrix}. \qquad [3]$$

If we refer to the off-diagonals of $\Lambda\Lambda'$ as validities and to the diagonals as reliabilities, in this case both A and B have a reliability equal to .36 and validity of .36. Now suppose Test A is administered under conditions that allow for partial knowledge and that, as a result, its correlation with the construct goes from .60 to .70. Following the same procedure, we now find that the reliability of A is .49 while that of B remains at .36, and that the correlation (validity) has gone up from .36 to .42. In short, when there is a common factor between two measures, an increase in the reliability of one of them will lead to an increase in validity. This is not so when more than one factor is common.

To illustrate this, assume that Tests A and B, both administered conventionally, have in common, in addition to the construct, a method factor, and that both correlate .40 with it. That is,

$$\underset{2\times 2}{\Lambda} = \begin{matrix} \theta_c & \theta_m & \\ \end{matrix}\begin{bmatrix} .60 & .40 \\ .60 & .40 \end{bmatrix}\begin{matrix} \text{Test} \\ \text{A} \\ \text{B} \end{matrix} \qquad [4]$$

Assuming that the construct and the method factor are uncorrelated in the population, the correlation matrix for A and B, according to the model in Equation 2, is given by

$$\Sigma = \overset{\Lambda}{\begin{bmatrix} .60 & .40 \\ .60 & .40 \end{bmatrix}} \quad \overset{\Lambda'}{\begin{bmatrix} .60 & .60 \\ .40 & .40 \end{bmatrix}} + \overset{\Psi}{\begin{bmatrix} .48 & .00 \\ .00 & .48 \end{bmatrix}}$$

$$= \overset{\Lambda\Lambda'}{\begin{bmatrix} .52 & .52 \\ .52 & .52 \end{bmatrix}} + \overset{\Psi}{\begin{bmatrix} .48 & 0.00 \\ 0.00 & .40 \end{bmatrix}}$$

$$= \begin{bmatrix} 1.00 & .52 \\ .52 & 1.00 \end{bmatrix}. \qquad [5]$$

The validity is .52.

Now suppose that Test A above is again administered under conditions that allow for the scoring of partial information and that, as a result of this, its correlation with the construct becomes .70. At the same time the correlation of Test A with the method factor drops from .40 to .20; i.e., Λ becomes

$$\Lambda = \begin{bmatrix} .70 & .20 \\ .60 & .40 \end{bmatrix} \quad \begin{matrix} \text{Test A (with partial knowledge)} \\ \text{Test B} \end{matrix} \qquad [6]$$

and

$$\Sigma = \begin{bmatrix} .53 & .50 \\ .50 & .52 \end{bmatrix}. \qquad [7]$$

Thus, as a result of introducing partial knowledge, the validity was reduced from .52 to .50. However, it is clear that this seemingly disappointing result is not inconsistent with the true improvement that occurred, namely an increase of the correlation with the construct.

Although this example contains many assumptions, it seems that something similar occurs with real data. Hakstian and Kansup (1975) compared the validity of a verbal ability test administered under conventional and elimination scoring (Coombs, Millholland & Womer, 1956) instructions. Validity was defined as the correlation with school grades in language arts. This correlation was .49 under

conventional administration and .39 under elimination scoring. However, the correlation with another verbal ability test was .59 under conventional scoring and .67 under elimination scoring. Thus, defining validity as the correlation with school grades, elimination scoring appears to be less valid; but defined as the correlation with another verbal ability score, elimination scoring is more valid. These two findings are not contradictory but simply provide evidence of the fact that school grades and test scores are not unidimensional.

Advantages of Using Partial Information

In short, I think a critical review of the literature will convince most that the question is not whether partial knowledge scoring improves the validity and reliability of test scores but rather under what conditions are gains to be expected, and how large those gains are likely to be, in particular whether they are large enough to offset any increase in testing time. It stands to reason that if methods for the assessment of partial knowledge are to yield improved test scores, the tests must be such that there will be an opportunity for partial knowledge to emerge. With few exceptions, most notably Coombs, *et al.*, the presence of partial knowledge is never tested. Some theoretical results suggest that when partial knowledge is allowed to emerge and it is scored, dramatic improvements in test scores follow.

To illustrate this, I computed the information functions of two latent trait models. (You will recall that information at a given point on the underlying trait is the reciprocal of the variance of the maximum likelihood estimate at that point. Therefore the larger the information value, the more precise our estimate of the location of an individual on the trait.) One of the models uses the two-parameter normal ogive which is appropriate for dichotomous scoring. The other model was Samejima's (1969) graded response model, which is an extension of the two-parameter normal ogive to polychotomous scoring. You may think of the information of the graded model as the case when partial knowledge is taken into account, whereas the information provided by the dichotomous model is that provided when partial information is ignored.

To simplify the comparison, I computed for each model the mean information assuming that the underlying trait was normally distributed. The ratio of the mean information for the graded model over that of the dichotomous model for several levels of test homogeneity is seen in Table 1. For example, at $r=.30$ the ratio is 1.42. This means that, on the average, the use of partial knowledge will be 42% more informative than if it is ignored. Note that this improvement, due to incorporating partial information into the scores, increases as the discrimination of the test increases. In other words, the better the test, the more it will benefit from adding partial knowledge.

Table 1
Ratio of Mean Information of Graded to
Dichotomous Model, as a Function of Inter-Item Correlation

	Inter-item correlation					
	.30	.40	.50	.60	.70	.90
Ratio of mean information	1.42	1.43	1.48	1.52	1.58	1.90

The advantages derived from taking partial knowledge into account can only materialize under the proper conditions. In the conventional testing situation, even though partial knowledge influences which alternative is chosen, the response is scored as correct or incorrect. One way of allowing credit to be given for partial knowledge is to instruct testees to segregate alternatives into different categories. Coomb's procedure is an instance of this approach where the categories are "correct" and "incorrect". Other categories are possible, though; for example, verbal items may be classified as synonyms, antonyms, or neither.

Computerized Testing

Recording and scoring responses to this kind of item is not, however, convenient with paper and pencil administration. This brings me to another aspect of our research, namely the use of computers. One obvious use of computers is to handle the recording and scoring of responses, but as previous presentations in this symposium suggest, the computer can also be used to adapt or tailor the test to each individual.

These presentations, and indeed most of the research in computerized adaptive testing, are oriented toward ability measurement. In achievement testing, we should distinguish between two kinds of tailoring. One is tailoring the length of the test and the other is tailoring the difficulty of the test.

Tailoring the length of the test is appropriate in instructional settings where each individual is allowed as much time as necessary to complete a given unit of instruction. Under those conditions, individual differences with respect to knowledge are minimized and it becomes profitable to tailor the test in terms of length rather than difficulty. The research of Ferguson (1970) is an example of this type of tailoring. In his system, an individual is tested until he is classified into a non-mastery or mastery category. The statistical basis of this system is that of Wald's sequential likelihood ratio test. Ferguson's model assumes that the difficulty and discrimination of all items are the same. It is not known how sensitive the procedure is with respect to violation of these assumptions. Research addressed to this question is needed. It would also be desirable to study the possibility of relaxing the model to allow for unequal item difficulties and discriminations as well as allowing for polychotomous responses.

Although self-paced instruction has many advantages, limited resources often do not permit its full implementation. As a result, the sample under instruction will likely be heterogeneous with respect to achievement. Similarly, if we are testing for retention of achievement or for levels of achievement acquired prior to instruction, we will also find wide variation in performance. Under these conditions, tailoring the test to an individual's level of achievement will be more efficient than the conventional non-adaptive procedure, as the previous presentations suggest.

One of the major aims of our research is to combine the advantages of partial knowledge scoring and adaptive testing. Most of the research on adaptive testing at the University of Minnesota and elsewhere has been done in the context of dichotomous response models. The exceptions are to be found in the work of Bayroff & Anderson (1960), Wood (1971) and Samejima (1975).

Bayroff & Anderson seem to be the only ones to have actually implemented an adaptive testing strategy using non-dichotomous items. Essentially what they did was to branch an individual according to the correctness of the alternative chosen. Although they used a polychotomous item for the first item only, this can be readily extended to include all items. Other branching rules are possible. Wood (1971) suggested that the optimal branching rule will administer as the next item the most discriminating of those items with a midpoint of adjacent categories closer to the individual's current estimate of achievement. Samejima (1975) carried out a simulation on live data of a similar procedure which she referred to as tailoring the dichotomization of the item to the individual. She noted dramatic improvements by comparing the plot of scores based on a uniform dichotomization and tailored dichotomization against the scores based on the polychotomous responses.

Summary

To summarize, one part of our research is concerned with the joint implementation of two recent developments in test theory: adapting the test to the individual and simultaneously extracting more information from each response by recording partial knowledge. The question that remains is whether sets of items can be constructed such that they will allow partial knowledge to be utilized without unduly increasing testing time. By next year's meeting, I hope to have the answer to this and other related questions.

REFERENCES

Bayroff, A.G., Thomas, J.J., & Anderson, A.A. Construction of an experimental sequential item test. Research Memorandum 60-1, Personnel Research Branch, Department of the Army, January 1960.

Coombs, C.H., Millholland, J.E., & Womer, F.B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 17-37.

Ferguson, R.L. A model for computer-assisted criterion-referenced measurement. Education, 1970, 91, 25-31.

Hakstian, A.R., & Kansup, W. A comparison of several methods of assessing partial knowledge in multiple choice tests: II testing procedures. Journal of Educational Measurement, 1975, 12, 231-240.

Samejima, F. Graded response model of the latent trait theory and tailored testing. Proceedings of the First Conference on Computerized Adaptive Testing. United States Civil Service Commission, Bureau of Policies and Standards, 1976.

Samejima, F. Estimating latent ability using a response pattern of graded responses. Psychometrika, 1969, Monograph Supplement No. 17.

Shoemaker, D.M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.

Wood, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation, University of Chicago, 1971.

Applications of Item Response Theory to the Problem of Test Bias

by
Steven M. Pine
University of Minnesota

Introduction

One of the most challenging and important issues facing test developers and users today is whether or not tests are biased against certain minority groups, and if so, what to do about it. In recent years, there has been a plethora of research activity to study and reduce bias and unfairness in various settings. For the most part, these efforts have been dismal failures. A careful review of research in this area indicates that one possible reason for this lack of progress is the fact that almost all the research on test bias and fairness has been carried out within the context of classical test theory.

In his recent review of test theory, Lumsden (1976) refers to the true score model of classical test theory as the "Model-T Theory" and suggests that classical test theory has failed as a theory. Therefore, it makes little sense to continue working within this antiquated framework to resolve problems as complex as those involved in the bias issue. Not only is classical test theory antiquated (and possibly useless as Lumsden suggests), but it emphasizes exactly the kind of group-oriented measurement that needs to be deemphasized if we can ever hope to make tests more relevant to individuals of varied backgrounds.

In essence, all bias in testing is caused by the failure of tests to take into account a number of important individual differences variables in their construction, administration, and scoring (Sattler, 1974; Angoff, 1975; Green, 1976; Pine & Weiss, 1976). These variables include individual differences in motivation, ethnic background and related variables.

Tests based on the classical model, because they are based on group statistics and because they require all testees to take identical test items, ignore individual differences. Therefore, trying to resolve the problems of test bias with the classical model is like trying to put a satellite into orbit with a submarine. What is needed if we are to make any progress in this critical research area is a theory that permits the testing process to be tailored to individuals. This technology now exists in the form of item response theory.

This research is supported by Contract N00014-76-C-0244, NR No. 150-383, with the Personnel and Training Research Programs, Office of Naval Research.

This paper will discuss some of the research I am conducting at the University of Minnesota, supported by the Personnel and Training Programs of the Office of Naval Research (NR 150-383), to apply item response theory to the problem of test bias. I will begin by defining bias in terms of item response theory and indicate how the definition leads to several testable implications. Next, I will demonstrate these concepts with data from a real set of test items. Just to whet your appetite, I will tell you now that these items turn out to be unbiased! Finally, I will briefly discuss some of the current research we are conducting on incorporating some of these ideas with computerized adaptive testing to examine their potential benefits in testing minorities.

An Item Response Model of Bias

Item response theory. Recently, a new test theory called "item response (or latent trait) theory", which was specifically designed for the measurement of individuals, has emerged. Briefly stated, item response theory is based on the idea that the response which an individual makes to a given test item is determined by his/her ability on one or more underlying dimensions (latent traits), and the parameters of the test items, i.e., their difficulty, discriminating power, and probability of being guessed correctly by chance. This idea is expressed mathematically by the Item Characteristic Curve (ICC) which gives the probability that a testee of a given ability level on the underlying dimension will answer the administered item correctly.

The ICC curves and their associated item parameters are the building blocks of item response theory. Once item parameters are determined for each test item, they completely describe how individuals at a given ability level will perform on each item. This allows probabilistic statements to be made about the ability of a testee regardless of subgroup membership or which subset of items has been administered. This property provides a means for creating tests which can be adapted to individual testees, since with this test model one is no longer required to administer identical items to every testee. It is this property that makes item response theory so potentially valuable for developing less biased tests. Furthermore, the bias-reducing potential of item response theory is not tied to its use with any particular testing strategy, although the greatest benefits can be expected when it is used in conjunction with adaptive testing.

I have claimed that item response theory can be useful in constructing less biased tests. I will now define exactly what I mean by "unbiased", state the assumptions that must be met, and indicate some testable implications of these assumptions.

Definition of item bias. A test item is unbiased if all individuals having the same underlying ability have an equal probability of getting the item correct, regardless of subgroup membership.

As has been previously mentioned, the ICC gives the probability of getting an item correct given an ability level. Therefore, the above definition of an unbiased item is equivalent to requiring that a test item have the same ICC for all subgroups. Since an ICC is determined by its parameters, this is also equivalent to requiring that the item parameters which describe an item's difficulty, discrimination, and probability of being guessed correctly by chance be invariant up to a linear transformation from subgroup to subgroup. (Linear transformation is necessary to account for the fact that subgroups in which the parameters are calibrated may have ability distributions with different means and variances.)

Assumptions and tests for unbiased items. In this paper, I will restrict my discussion to tests that consist entirely of homogeneous items. By homogeneous, I mean items which measure approximately one ability dimension. The reason I say approximately is that I am allowing for the possibility that these items may measure one or more extraneous dimensions as well. For instance, test items intended to measure vocabulary ability may inadvertently also measure several cultural variables. By restricting ourselves to homogeneous items, we will have an easier time developing a theoretical foundation that can later be extended to the multidimensional case.

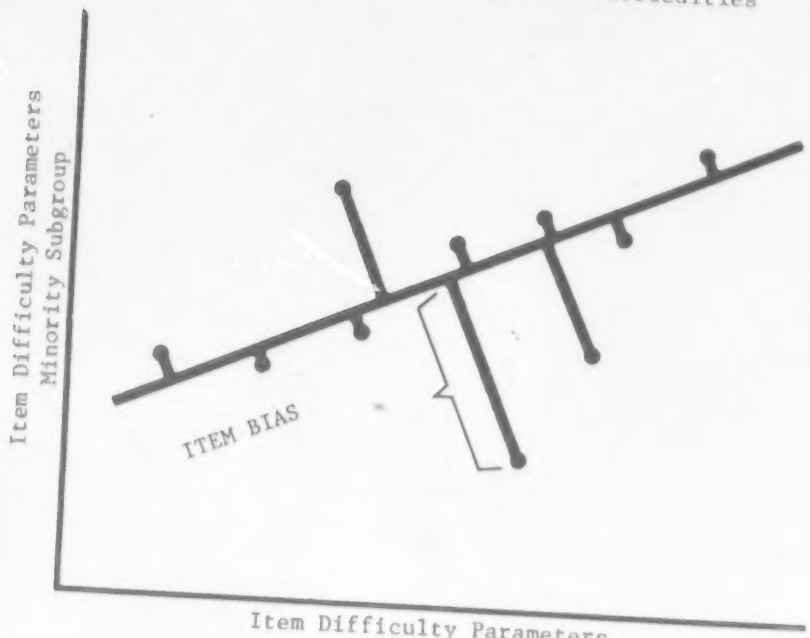
Lord & Novick (1968, p. 359) have shown that the parameters of a test item will be invariant across subgroups if the same dimension governs performance on that item for each of these subgroups. According to the definition given earlier, this implies that test items are unbiased if they measure the same single trait for all subgroups. This fact suggests that the bias in a test can be directly assessed by factor analyzing the matrix of inter-item correlation coefficients for each subgroup in the population and demonstrating that the non-random variance in each matrix is accounted for by a single factor.

Furthermore, if the same dimension underlies a set of test items for a population of testees (which would, therefore, make the items unbiased), the item parameters for any two subgroups in the population ought to have a linear relationship (Lord & Novick, 1968, p. 380). This condition can be tested directly by plotting the discrimination (a), difficulty (b), and guessing (c) parameters of one subgroup against another and testing for linearity. A plot of this type is shown in Figure 1.

Also, by comparing the relative item parameter values between subgroups, the specific test items which contribute the most to causing the item by group interaction to be nonlinear can be identified. In Figure 1, the perpendicular distance between each item and the best fitting line through all the points can be interpreted as the degree of item bias. Similar plots and interpretations could also be made for the item discrimination and guessing parameters.

The degree-of-item-bias index could be used to screen out the most biased items during the construction of a conventional test, or could be used within an adaptive testing framework as an additional item selection criterion.

Figure 1
Item Bias Shown as a Perpendicular Distance
in a Scatter Plot of Subgroup Item Difficulties



The procedure of assessing item bias by plotting a scatter diagram of item parameters for one subgroup against another is not in itself new. A very similar method has been used at Educational Testing Service (ETS) for years. The essential difference between the present method and the ETS method is that the ETS method uses item parameters based on classical test theory. It can be shown (Lord & Novick, 1968, p. 381) that classical item parameters will not, in general, be linearly related across subgroups of a population. This means that the test for bias using classical parameters can lead to an artifactual detection of bias. Furthermore, the difficulty parameter of classical test theory is confounded by level of discrimination and guessing effects (Urry, 1974). Therefore, if an item is relatively more difficult for one subgroup than another, one really does not know whether this is because the item varies on difficulty, discrimination, guessing, or some combination thereof.

The item response parameters, on the other hand, provide relatively unconfounded measures of difficulty, discrimination, and guessing. Therefore, by plotting these parameters on separate plots, one can determine exactly why an item is biased. For instance, for any given item it may turn out that an item is biased not because it is relatively more difficult for a minority subgroup, but because this subgroup is less effective at guessing. This kind of detailed analysis is impossible using classical item parameters.

An example with real data. To demonstrate how these analyses might be used and interpreted, I have applied them to 75 multiple-choice vocabulary items collected at a racially mixed high school in Minneapolis. The sample sizes in this study were smaller (58 Blacks, 168 Whites) than we would normally like, but the data provide a good example of the technique.

First, we can test the unidimensionality assumption by factor analyzing the inter-item correlation matrices. A subset of 45 items was chosen and two tetrachoric correlation matrices were calculated on these items, one for the Black and one for the White subsamples. The matrices were then factor analyzed using the principal axis method. Eight factors were extracted from each matrix. The eigenvalues from the two factor analyses are shown in Table 1.

Table 1
Eigenvalues from Factor Analysis of Black and White Subgroup Item-Intercorrelation Matrices

Subgroup	Factor	Eigenvalue	Percent of Variance	Cumulative Percent
Whites				
	1	19.26	64.8	64.8
	2	2.32	7.8	72.7
	3	1.67	5.6	78.3
	4	1.58	5.3	83.7
	5	1.37	4.6	88.3
	6	1.20	4.1	92.4
	7	1.18	4.0	96.4
	8	1.08	3.6	100.0
Blacks				
	1	16.33	47.9	47.9
	2	3.70	10.9	58.7
	3	3.01	8.8	67.5
	4	2.64	7.7	75.3
	5	2.35	6.9	82.2
	6	2.26	6.6	88.8
	7	2.06	6.0	94.9
	8	1.75	5.1	100.0

For both the Black and the White data, the first eigenvalue was very large and, therefore, the unidimensionality assumption seems to hold. Furthermore, the items appear to be measuring the

One of the first things to notice is that almost all of the items are relatively more difficult for Blacks than for Whites. This is indicated by the fact that the items tend to fall below the 45° angle line. The points would fall on this line if they were equally difficult for Blacks and Whites. However, in this analysis, the mere fact that the items are relatively more difficult for Blacks cannot necessarily be taken as an indication of bias. Bias in the test items is assessed by evaluating the degree of linearity in the plot.

For these items the relationship appears to be quite linear and, in fact, the correlation coefficient between Blacks and Whites is r=.86. This is consistent with the results of the factor analysis and indicates that these vocabulary items are essentially unbiased.

You may be curious to know what an unbiased set of test items looks like. I have, therefore, appended a sample of six of these items to this paper for your perusal. You will notice that there is nothing particularly special about any of these items. The items have not been specifically chosen to reflect minority cultures nor do they employ fancy response options. That is not to say that these methods are necessarily undesirable; the point, however, is that the extent to which items are biased can only be determined by an empirical investigation of how the items behave in each of the subgroups of the population.

If a more in-depth analysis of these items were desired, one could produce and interpret scatter plots for the item discrimination and guessing parameters in a similar manner. This will not be done in this case because reliable estimates of these parameters cannot be obtained with the sample sizes that were available. Another analysis that could be performed on the plot in Figure 1 is to determine the index of bias for each item. It is possible that even though the items taken as a group are unbiased, one or more of the items taken individually might be biased. For instance, in this data set, several items appear to have unusually large departures from linearity. Of course, it is possible that these large departures may be due only to sampling error. To eliminate possible misinterpretations that would occur if this were the case, I am working on a technique to establish confidence limits for the best fitting line so that we can say with some known degree of confidence that any item which exceeds these limits is biased.

Concluding Remarks

The material presented here is only a small example of how item response theory can potentially be applied to the problem of test bias. In fact, it represents only a small part of the research currently underway at the University of Minnesota. Nonetheless, I believe that this material provides a simple, but cogent example of the efficacy of applying item response theory to the bias problem.

In the time which remains, I would like to briefly describe some of our other research projects. The first project is really an extension of the research described here. It involves a method of correcting for bias in the item parameters. Very briefly, this method consists of determining item parameter estimates so that they will depend only on the extent to which an item loads on the factor it is supposed to be measuring.

In essence, this approach is based on the notion that to obtain unbiased test items, it is not necessary to expend huge amounts of time and money searching for so-called "culturally-relevant" test items. It does not really matter what the test items look like. All that is necessary is to know how each test item behaves (i.e., what its parameters are) in the various subgroups which comprise our test population. If we know that an item is biased, we can correct its parameter values, e.g., with the method I have suggested, to account for this bias. Then, if we base our test scores not on the total number of correct answers, but on some function of the corrected parameter values, we can produce unbiased ability estimates.

We are now in the process of checking out this method for correcting item bias with computer simulations. In this way, the estimated corrected-item-parameter values can be directly compared to the known, true-item-parameter values. If the results of the study are favorable, we will have devised a means of eliminating the effect of bias in test items.

Does this mean that we can now write the final chapter on test unfairness? No, not at all. First, I am not so naive as to think that a solution to the bias problem will ever be universally accepted as long as differences exist in the mean test scores of various subgroups. Secondly, bias in the estimation of item parameters is only one source of possible unfairness in the test process. A test can be unfair for a myriad of other reasons, including those attributable to elements in the testing environment, and to the procedures used to select and administer test items.

To explore the possible psychometric influences on unfairness, we have been conducting a series of computer simulations to investigate how the item characteristics of a test interact with the choice of a testing strategy, to influence test fairness in selection.

Finally, I would like to mention the research we are currently doing in the Minneapolis Public School System to show that I test people as well as electrons. In this study, we are investigating just how well some of the procedures mentioned here hold up in a real test administration. We will also be using the study as a proving ground for our recently developed bias-reducing computerized adaptive test, and to replicate a previous finding that computerized tests increase the test-taking motivation of minority testees. Perhaps next year a report of the results of these research efforts will be possible.

Appendix: A Sample of Six Unbiased Test Items

rudimentary	1	impolite
	2	uncivilized
	3	elementary
	4	protected
	5	destructive
ecstatic	1	dotted
	2	fawning
	3	proper
	4	inept
	5	sullen
abacus	1	computing device
	2	flower
	3	herb
	4	musical instrument
	5	carpentry tool
gnome	1	dwarf
	2	lizard
	3	ghost
	4	rock
	5	miner
capricious	1	frugal
	2	callous
	3	medicinal
	4	tropical
	5	whimsical
exemplary	1	laudable
	2	inside
	3	quick
	4	lithe
	5	magic

REFERENCES

Angoff, W.H. The investigation of test bias in the absence of an outside criterion. Paper presented at the NIE Conference on Test Bias, December 1975.

Green, D.R. Reducing bias in achievement tests. Paper presented at the NIE Conference on Test Bias, San Francisco, April 1976.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Lumsden, J. Test theory. In Rosenzweig, M.R., & Porter, L.W. (Eds.), Annual review of psychology. Palo Alto, CA: Annual Reviews, Inc., 1976.

Pine, S.M., & Weiss, D.J. Effects of item characteristics on test fairness. (Research Report 76-5). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1976.

Rummel, R.J. Applied factor analysis. Evanston, IL: Northwestern University Press, 1970.

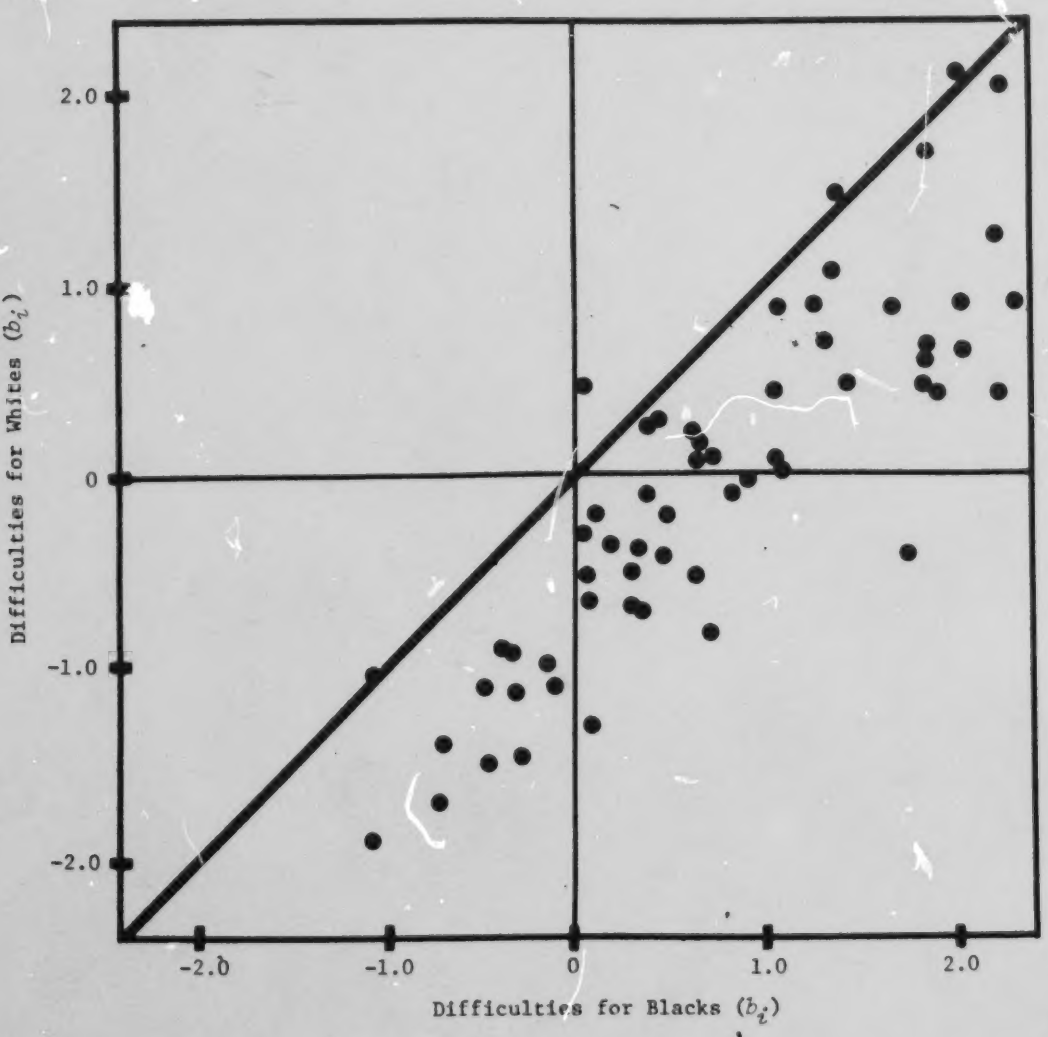
Sattler, J.M. Assessment of children's intelligence. Philadelphia: W.B. Saunders Company, 1974.

Urry, V.W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

same dimension in both subgroups, because the coefficient of congruence (Rummel, 1970, p. 461) calculated between the 45 corresponding loadings for Factor I in the two subgroups was .97.

Now let's look at the results of the bias analysis carried out on all 75 items. The scatter plot in Figure 2 shows the result of plotting each item by its estimated difficulty parameter in the White subsample versus its estimated value in the Black subsample.

Figure 2
Race Contrasts of Item Difficulties for 75 Vocabulary Items



AFHRL Research on Utilization of Women in the Air Force

by
Jeffrey E. Kantor
Personnel Research Division
Air Force Human Resources Laboratory

The military environment is constantly subject to change resulting from both external and internal forces. Currently modifying the military environment are: reductions in financial allocations; reductions in authorized personnel strengths; increased emphasis on compliance to federal directives on equal opportunity employment; changes in service policies, such as the admission of women to flight training; and the trend toward increasing the numbers of women in military service. These influences have combined to produce certain general and specific managerial problems not recently, or in some cases, ever previously encountered. In general, it has become crucial to insure that all military positions are filled with the best person for the job; whether male or female. Consequently, women, in substantial numbers, are now entering career fields traditionally staffed by male personnel. Specific questions then arise: can screening, selection and assignment procedures be equally applied and be equally effective for both sexes, do similar factors affect the training and on-the-job performance of both sexes, and do any male/female differences exist which mandate consideration with regard to personnel decisions?

The Air Force Human Resources Laboratory is currently conducting a program of research efforts designed to obtain information needed to answer these type questions and ready the Air Force to meet the changing demands of

the current and future military environment. For presentation purposes, I shall categorize these efforts into ongoing research incorporating some male/female comparative aspects and research initiated specifically to delineate male/female comparisons. Additionally, I should like to add that these research efforts are far from complete, and the data which I am able to report now represent only fragments of the anticipated products. Therefore, please consider this presentation as an overview of what we are doing rather than a synopsis of what we have done.

Many of the individuals who leave service prior to the normal termination of their first term are those who have experienced problems in adapting to military life. Identification of these personnel before enlistment is the goal of suitability screening. Previous research, on a sample of male, draft motivated enlisted accessions (Guinn, et. al., 1975) has led to the conclusion that the History Opinion Inventory (HOI), a 100-item objective questionnaire, has potential usefulness as a screening device for suitability to military service. Current research is evaluating the usefulness of this instrument on a 1975 sample of 32,000 male and 8,000 female subjects. Initial analyses appear to indicate that the HOI could be useful and applicable for both male and female enlisted accessions. These preliminary analyses are based on the criterion of graduation vs elimination in basic military training. As technical training and job performance criteria become available, further analyses will be accomplished. The final objective of this project is to develop an instrument which allows the selection of the most potentially suitable male and female enlisted personnel.

To increase the precision and efficiency of the classification and assignment process, other information, in addition to traditional aptitude

data, is needed. Under a contract from the Human Resources Laboratory, Educational Testing Service has developed an interest inventory, the Vocational Interest Career Examination (VOICE), specially tailored to military needs. To insure that the VOICE is an effective tool in the assessment of interest patterns, research on the predictive validity of this instrument is now underway. Ten thousand male and 10,000 female enlisted accessions, scheduled for assignment in one of twenty DoD occupational areas, were administered the VOICE during basic military training. A one-year follow-up of these personnel has provided a criterion of expressed job satisfaction. Analyses conducted so far have indicated: male and female norms for the various scales are in some cases slightly different; however, no differences were found between male and female response patterns, i.e., inter-item correlations and factors structures were the same for male and female respondents. Additionally, using identical regression equations, the VOICE was found to be equally effective, in predicting job satisfaction for both males and females. An interesting spin-off of this research is the comparison of expressed job satisfaction between male and female respondents in similar jobs. In general, males and females reported the same degree of satisfaction with two exceptions, these being that males reported more job satisfaction in both the general aircraft and utilities maintenance fields (Alley, et. al., 1976a, 1976b). The final objective of this research effort is to provide an instrument to assist recruiters in their role as vocational counselors. This may prove to be of particular benefit regarding the assignment of women to non-traditional career fields.

Technical training of Air Force personnel represents a considerable financial and manpower investment, and resources are continually being lost through attrition in training. Research is now in progress to probe the factors, other than aptitude, which may be related to performance in the technical training environment. Specifically, an instrument was constructed to tap student attitudes about motivation, classwork, instructors, fellow students, physical aspects of the training environment, and satisfaction with the training experience. This "Technical Training Student Survey" was administered to 11,000 male and 1,800 female students attending five Air Force technical training centers. Analyses are just in the preliminary stage but appear to support at least the following conclusions: the instrument has suitable psychometric properties; the instrument is sensitive to student attitudes; and there are significant differences between graduating and non-graduating subjects. In particular, graduates reported that performing well was more important to them than it was to non-graduates; graduates saw the instructors as being more competent and personable, and graduates were more satisfied with both their career choice and overall environmental situation than were the non-graduates. Additionally, non-graduates reported more organizational control and stress than did the graduates. Preliminary findings include some male/female differences of interest. Females reported slightly more organizational control and training stress; less satisfaction with the physical environment; and a greater degree of inaccuracy of pre-service information about military life. Perhaps these differences might be summarized as females being as highly motivated as males, but slightly less satisfied with the overall training environment. A substantial amount of work remains to be accomplished on this project and should

produce much more male/female comparative data.

Increasing numbers of women are now being assigned to Air Force Security Police career fields. There are actually two separate security police career fields in the Air Force. The Law Enforcement (812XX) career field is rather analogous to civilian police, while the Security Police (811XX) career field is relied upon for equipment, munitions, personnel, and air base security and defense. Women have been assigned to the law enforcement field for several years and just recently are being assigned to the security police (811) field. Currently; we are studying the factors related to on-the-job performance and attrition in both fields. Two data gathering approaches have been taken on this project. First, a group of 4,300 male and 300 female trainees were administered a battery of tests while in technical training. The careers of these personnel are now being monitored to determine the predictive usefulness of the measures which included bio/demographic, personal interest and opinion information. Few data are presently available from this approach. Second, in this approach, the same battery of tests was administered to 1,800 male and 120 female security police already on the job. Supervisor ratings of job performance were also obtained on these people. One interesting finding already available is that there were no differences in overall job performance ratings between males and females. This is somewhat contradictory to anecdotal reports previously available. Additionally, similar positive and negative perceptions of the job environment were obtained from members of both sexes. Future analyses are expected to reveal pertinent information on the performance of male and female members of security police under stress situations.

Recently, the Human Resources Laboratory has begun several research projects specifically designed to identify and assess possible male/female differences and similarities. One area of research involves psychomotor performance. Psychomotor ability has been found to be correlated with performance in several career fields, such as mechanics, electronics, and of course, flying. Using newly developed electronic equipment, we are now studying the psychomotor performance of both enlisted and officer females on a wide range of tasks. These data will be compared to male data already on hand and assessed for similarity in terms of performance and predictive usefulness in terms of later on-the-job success. Indications presently available appear to show that there are little or no differences in the psychomotor abilities of males and females.

In August 1976, the Air Force began training its first class of female pilots. Prior to training, we were able to extensively test these women on a variety of psychomotor and other tasks found to be associated with success in pilot training. Values obtained from these subjects compared quite favorably in all respects to similar data from previously tested male subjects. Research in the psychomotor area for female subjects is just getting underway and is anticipated to produce fruitful results over the next few years.

Two contract research efforts are also directly related to the utilization of women in the Air Force. One contract will investigate the civilian utilization of women in industrial career fields. The technical objectives of this project are to: identify civilian managerial policies as they relate to the employment/non-employment, retention/dismissal, and advancement of women employees working in production type jobs. Additionally to assess the differences and similarities of civilian and military policies; and project

the impact of the possible adoption of civilian policies by the Air Force. To obtain information relevant to these issues, corporate officials at major US industrial concerns will be interviewed and surveyed. This effort is just being established and is, at present, still in the negotiation stage. It is hoped that by evaluating civilian policies and experience, it may be possible for the Air Force to adopt maximally effective managerial policies without incurring the costs associated with a trial and error process.

The second contract will evaluate male/female differences in civilian aviation. The technical objectives of this project are to: investigate aircraft accidents and determine if either sex appears disposed towards a greater percentage of accidents or types of accidents; identify and assess differences in instructional needs, student performance, and potential; and identify and assess motivation and interest in civilian and military flying. The data will be obtained through several sources including survey administration to FAA certified flight instructors and pilots, and close investigation of federal accident report data. Findings from this project will be evaluated in terms of impact on Air Force flight training and proficiency programs. This contract is also in the negotiation stage.

In summary: the Air Force Human Resources Laboratory is attempting to investigate both the general and specific issues related to the maximal utilization of women in the Air Force. These projects are far from complete, but an evaluation of the available data suggests some conclusions.

First, there are some significant attitudinal differences between males and females in specified areas, but none that would appear to negatively impact on most military assignments. Second, factors which either negatively or positively affect members of one sex similarly affect members of the other sex.

(525)

Finally, techniques and procedures developed for the selection, training, assignment, and utilization of one sex appear to work equally well with the other sex. These findings are still preliminary with much more research to be done in the future. However, it does appear that the female segment of the civilian population can be extensively tapped to provide many of the personnel resources necessary to meet the requirements of the current and future military environment.

References

Alley, W. E., Berberich, G. L., and Wilbourn, J. M. Dimensions of vocational interest. AFHRL-TR-76-00. Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory, 1976a (in press).

Alley, W. E., Wilbourn, J. M., & Berberich G. L. Predictive relationships between occupational interests and reported satisfaction in vocational and technical career fields. AFHRL-TR-76-00. Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory, 1976b (in press).

Guinn, N., Johnson, A. L., & Kantor, J. E. Screening for adaptability to military service. AFHRL-TR-75-30. Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory, May 1975.

THE UTILIZATION OF WOMEN IN THE MARINE CORPS

STEVEN GORMAN
HEADQUARTERS MARINE CORPS
WASHINGTON, D.C.

STEVEN GORMAN
U. S. MARINE CORPS

This presentation will give a brief overview of the utilization of women within the Marine Corps. Included will be data on women working in various occupational fields. This information will show a lower percentage of women in the traditional clerical field and greater utilization in nontraditional occupational fields. This trend parallels changing societal views on women in careers. In addition, data gathered will show a greater percentage of women utilized in both enlisted and officer structures. Additionally, attitudes of women within this service will be examined, as well as the attitudes of men toward women in the service.

QUALIFICATIONS AND QUALITY FOR WOMEN

Women enlisted in the Marine Corps must be high school graduates or General Educational Development (GED) equivalent. In comparison, only 67.5% of FY76 male accessions were high school graduates or GED's. The average AFQT-type mental ability score for FY76 accessions was 76.6 for women and only 59.8 for men. Of the female accessions, 11.5% had some education past high school, and 8.3% had Bachelor degrees. For male accessions, only 4.0% had any education past high school. Additionally, the reenlistment rate and retention rate for first-timers and careerists was higher for women than for men. In general, women are superior to men on these quality indicators of enlisted accessions.

Requirements for officers are the same regardless of sex. These are: (1) graduation from an accredited four-year degree granting institution, and (2) an appropriately high score on an ability test, the Officer Aptitude Rating.

An interesting set of demographic data shows that most male officers are married (81%) while most female officers are single (66%).

Women are represented equally throughout the enlisted hierarchy. See Chart A. Additionally women officers are represented up through the rank of Colonel. See Chart B.

CHANGES IN ACCESSION GOALS

In 1972, the Corps increased its enlisted end strength goals for women from 2,050 at the end of FY72 to 2,330 at the end of FY75.

As the table shows, actual end strength was 2,841 at the end of FY75 and 3,063 at the end of FY76. During the same period, end strength for women officers rose from 282 in FY72 to 366 at the end of FY76.

WOMEN MARINE STRENGTH LEVEL

	OFFICER	ENLISTED
FY71	277	1,981
FY72	282	2,066
FY73	315	1,973

FY74	336	2,402
FY75	345	2,841
FY76	386	3,003

In June 1976, the following decisions were made on women accessions.

a. Increase women officer accessions from 60 to 75 per year beginning FY77.

b. Increase annual enlisted accessions in accordance with the following schedule:

FY77	1,700
FY78-80	2,500
FY81-82	3,200

c. The projected Women Marine force will increase as follows:

	OFFICER	ENLISTED
FY77	429	3,484
FY78	470	3,735
FY79	500	4,619
FY80	528	5,290

RESTRICTIONS ON WOMEN ASSIGNMENT

The following restrictions apply to Women Marine assignment:

a. Title 10, USC, Section 6015

Not in aircraft engaged in combat missions nor on Navy vessels except hospital ships and transports.

b. National Policy

Not in a combatant role.

c. Manpower Control and Utilization Manual (MCO P5311.6A)

"...any non-FMF activity and the two FMF Headquarters. With the approval of HQMC, Women Marines may also be assigned to CONUS Division, Aircraft Wing, Force Troops and Force Service Regiment headquarters staffs, and subordinate units in rear echelon billets requiring service support, aviation support or communications specialties which will not require them to deploy with the assault echelon of the command if a contingency arises."

Additionally, certain occupational fields have been limited to women in the past. In 1972, the Marine Corps initiated a program to increase the number of occupational specialties open to women. The impact of this program is noted in Chart C. Several additional occupational fields are represented with women while at the same time, the percentage of women filling traditional areas, such as the 01 Occupation Field "Personnel and Administration" has decreased.

The Marine Corps conducted a study to reevaluate all occupational specialties during March 1975. The recommendations, approved in July 1975, provided for the aforementioned increases in the Women Marine objective force. All fields are open to women, with the exception of Infantry, Artillery, Tank and Amphibian Tractor and Pilot/Flight Crew occupational fields. All other fields are open to women subject to certain management considerations. The number of women to be classified in each military occupational specialty (MOS) is limited by such factors as preservation of the male overseas rotation base, equal job opportunity and advancement for all Marines, and the availability of non-deployable billets assignable to women.

A pilot program to assign women to the headquarters of CONUS Fleet Marine Force (FMF) commands was implemented in February through May 1974. On the basis of the effectiveness of that program, in February 1975, the routine assignment of women to CONUS FMF commands within the limitations of the maximum number by grade and skill recommended by the FMF commanders commenced. The assignment of Women Marines to the FMF would be not only at the headquarters level, but also to subordinate units which would not require them to deploy with the assault echelon of the command if a contingency should arise.

ADDITIONAL STUDIES AND IMPLICATIONS

At the present time, a study is being conducted to determine what additional CONUS as well as overseas FMF billets could be utilized by women without decreasing the command's deployability. The results of this study will be forthcoming this month, at which time further Women Marine assignment decisions may be made.

The Marine Corps commissioned the California State University at Los Angeles to review existing techniques and develop management tools necessary to conduct a thorough task analysis of all MOS's. This study will include the development of physical and operational standards where either is a factor in effective performance. The final study report has not yet been submitted to the Marine Corps.

Ongoing studies are conducted by the Systems Unit of the Manpower Planning, Programming, and Budgeting Branch (MPP) to determine the maximum feasible number of Women Marines that can be assigned to the various MOS's.

The results of these studies will provide additional non-traditional billets for Women Marines. The overall impact on the Marine Corps is an increase in the quality of the enlisted force, and a lowered pressure on recruiting new personnel, since quality women applicants are readily available and tend to stay in the military longer, thus lowering future quotas. The main interim problem is the lack of housing and training facilities, however, expansionary construction has been budgeted.

REFERENCES

Job Opportunities for Women in the Military: Progress and Problems, Comptroller General of the United States Report B-157371, May 11, 1976.

Service Market Share by State, Manpower Research and Data Analysis Center Report 76/6, Alexandria, Virginia, August, 1976

Thomas, Patricia J. (1976) Utilization of Enlisted Women in the Military, Navy Personnel Research and Development Center Technical Note 76-7, San Diego, California.

CHART A

TOTAL ENLISTED STRENGTH

AS OF 30 JUNE 76

	MALE	FEMALE	TOTAL
SGT MAJ / MGY SGT	1,204	6	1,210
1ST SGT / MSGT	3,275	20	3,295
GY SGT	8,653	53	8,706
SSGT	14,063	180	14,243
SGT	23,463	465	23,928
CPL	28,913	624	29,537
LCPL	35,457	721	36,178
PFC	32,257	626	32,883
PVT	23,169	368	23,537
TOTAL	170,454	3,063	173,517

CHART B
TOTAL OFFICER STRENGTH
AS OF 30 JUNE 76

	MALE	FEMALE	TOTAL
GENERAL	69		69
COLONEL	589	6	595
LIEUTENANT COLONEL	1434	14	1447
MAJOR	2920	39	2959
CAPTAIN	4754	96	4850
FIRST LIEUTENANT	4307	115	4422
SECOND LIEUTENANT	3262	99	3361
WARRANT OFFICER	1099	17	1116
TOTAL	18,433	386	18,819

ENLISTED WOMEN BY OCCUPATIONAL FIELDS

OCCUPATIONAL FIELD		31 DEC 1972		1 SEP 1976	
		NUMBER IN OCC FLD	PERCENT IN OCC FLD	NUMBER IN OCC FLD	PERCENT IN OCC FLD
01	Personnel and Administration	740	37.34	946	30.99
02	Intelligence	6	.30	2	.07
04	Logistics	28	1.41	17	.56
11	Utilities	-	-	31	1.02
13	Engineer, Construction, Equipment and Shore Party	-	-	27	.88
14	Drafting, Surveying and Mapping	-	-	7	.23
15	Printing and Reproduction	5	.25	16	.52
21	Ordnance	3	.15	2	.07
25	Operational Communications	124	6.26	161	5.27
28	Telecommunications Maintenance	48	2.42	154	5.04
30	Supply Administration and Operations	315	15.89	453	14.84
31	Transportation	37	1.87	66	2.16
32	Repair Services	-	-	9	.29
33	Food Service	44	2.22	16	.52
34	Auditing, Finance and Accounting	99	4.99	141	4.62
35	Motor Transport	-	-	157	5.14
40	Data Systems	92	4.64	103	3.37
41	Marine Corps Exchange	39	1.97	71	2.33
43	Public Affairs	29	1.46	54	1.77
44	Legal Services	15	.76	70	2.29
46	Photography	16	.81	21	.69
49	Training and Training Aids	27	1.36	48	1.57
55	Band	-	-	26	.85
58	Military Police and Corrections	-	-	86	2.82
59	Electronic Maintenance	-	-	1	.03
60	Aircraft Maintenance	-	-	3	.10
66	Avionics	28	1.41	53	1.74

CHART C

ENLISTED WOMEN BY OCCUPATIONAL FIELDS (CONT.)

OCCUPATIONAL FIELD	31 DEC 1972 NUMBER IN OCC FLD	31 DEC 1972 PERCENT IN OCC FLD	1 SEP 1976 NUMBER IN OCC FLD	1 SEP 1976 PERCENT IN OCC FLD
67 Air Control	23	1.16	-	-
68 Weather Service	28	1.41	23	.75
70 Aviation Operations	26	1.31	24	.79
73 Air Traffic Control	-	-	16	.32
98 Band of the USMC	-	-	18	.59
99 Basic Marine	200	10.09	231	7.57
Unknown	10	.50	-	-
TOTAL	1,982		3,063	

539

597

2

598

WOMEN MARINE OFFICERS BY OCCUPATIONAL FIELD

OCCUPATIONAL FIELD		NUMBER OF TOTAL WOMEN IN OCC FLD	PERCENT OF TOTAL WOMEN IN OCC FLD
01	Personnel and Administration	131	33.50
02	Intelligence	16	4.09
13	Engineer, Construction, Equipment and Shore Party	1	.26
25	Operational Communications	19	4.86
30	Supply Administration and Operations	26	6.65
31	Transportation	5	1.28
33	Food Service	8	2.05
34	Auditing, Finance and Accounting	37	9.46
35	Motor Transport	6	1.53
40	Data Systems	17	4.35
41	Marine Corps Exchange	11	2.81
43	Public Affairs	12	3.07
44	Legal Services	23	5.88
46	Photography	5	1.28
49	Training and Training Aids	3	.77
58	Military Policy and Corrections	5	1.28
60	Aircraft Maintenance	4	1.02
68	Weather Service	3	.77
70	Aviation Operations	1	.26
72	Air Controller/Air Support/ Anti-Air Warfare	4	1.02
73	Air Traffic Control	14	3.58
99	Basic Marine	40	10.23
	TOTAL	391	

PERCENT OF WOMEN IN TOTAL OFFICER FORCE 5.57

RESEARCH ON THE AERONAUTICAL ADAPTABILITY OF WOMEN

Annette G. Baisden

Naval Aerospace Medical Research Laboratory
Pensacola, Florida 32508

ABSTRACT

Although women have had some visibility in aviation for many decades, their participation beyond clerical, stewardess or the private pilot level is still perceived as novel. Military aviation has been opened to women on a basis of career parity with men. The motivation behind this is twofold: 1) response to the influences of current cultural and legal demands for equality of opportunity for women, and 2) the prospect of an enlarged person-power base from which future personnel needs of an all-volunteer force can be filled. This expanded opportunity placed a burden on the behavioral and life sciences to: (1) examine the role of women in aviation, (2) take stock of existing knowledge, (3) define areas of true differences or commonality with respect to the existing male aviation population, (4) identify problem areas, and (5) seek solutions. This overview will attempt to integrate certain existing knowledge and current Navy research concerning women in aviation. The specific topics to be addressed cover attitude, aptitude, performance, and human factors considerations.

The anthropometric problems related to the aeronautical adaptability of women which are discussed include: (1) design requirements for flight safety, emergency egress and operator comfort, and (2) strength requirements related to aircraft flight control.

Other research efforts discussed are (1) the measurement of attitudes of Navy and Marine Corps aviation personnel toward women on several factors, including their acceptance into aviation training, (2) the assessment of the female's ability to handle multiple complex tasks under stress, and (3) the effects of the menstrual cycle on complex perceptual-psychomotor tasks.

RESEARCH ON THE AERONAUTICAL ADAPTABILITY OF WOMEN

Annette G. Baisden

Naval Aerospace Medical Research Laboratory
Pensacola, Florida 32508

Although women have had some visibility in aviation for many decades, their participation beyond clerical, stewardess or the private pilot level is still perceived as novel. Military aviation has been opened to women on a basis of career parity with men. The motivation behind this is twofold: 1) response to the influences of current cultural and legal demands for equality of opportunity for women, and 2) the prospect of an enlarged person-power base from which future personnel needs of an all-volunteer force can be filled. This expanded opportunity placed a burden on the behavioral and life sciences to: (1) examine the role of women in aviation, (2) take stock of existing knowledge, (3) define areas of true differences or commonality with respect to the existing male aviation population, (4) identify problem areas, and (5) seek solutions. This overview will attempt to integrate certain existing knowledge and current Navy research concerning women in aviation. The specific topics to be addressed cover attitude, aptitude, performance, and human factors considerations.

The area of anthropometry in aviation has not received significant attention until recently and only in the last five years have the problems caused by mismatch between cockpit size and physical dimension of the pilot been recognized outside the anthropometry community. These problems can be considered

under three general headings -- design requirements for flight safety, emergency egress, and operator comfort (Lane, 1974).

Military aircraft are designed according to a set of specifications or standards which currently require that the cockpits accommodate 95% of the population of operators -- in this case, male military aviators. Most bodily dimensions of women are so much smaller than men that they represent a completely distinct population. Table 1 shows the differences between a population of male naval aviators and a civilian female population on six anthropometric dimensions. The last four dimensions are most critical in the design of aircraft cockpits. As shown in the table, over 95% of the male population exceed the female median on most dimensions. Women are also likely to encounter problems with the ejection seat which is designed to function optimally only for certain ranges of center-of-gravity. Although little information is available, it is possible that a large proportion of women will fall outside the design limits for ejection systems because of differences in height and weight distribution. In reference to operator discomfort and inconvenience, protective equipment and restraint systems were designed exclusively for the male aviator whose proportions in general do not correspond to those of women

Table 1

Comparative Data Indicating Male-Female Anthropometric Differences

ANTHROPOMETRIC DIMENSION	APPROXIMATE MEAN DIFFERENCE (INCHES)	APPROXIMATE PERCENTAGE OF MALES EXCEEDING FEMALE MEDIAN
WEIGHT	40 - 45 (lbs)	99.5
HEIGHT	5.5 - 6.0	99
SITTING HEIGHT	2.2	96
ARM REACH (estimated)	4.0 - 5.0	99.7
BUTTOCK-KNEE LENGTH	1.7	96
LEG LENGTH	1.5 - 2.0	85

The problems faced by women in the cockpit are by no means unique. The small and large males have long been a disenfranchised minority in cockpit design. From a realistic point of view only minimal changes are possible in aircraft design. However, significant progress can be made by redesigning protective equipment and by developing and fully implementing assignability codes to restrict aviators from aircraft they do not fit.

To date, the literature contains little objective data for either males or females to substantiate or correlate body strength requirements with strength requirements relative to aircraft flight control. Currently, studies are underway which should provide scientific, objective data necessary in establishing screening criteria for physical strength required of male and female applicants to flight training. Additionally, a direct comparison of strengths measured in closely simulated cockpit environments will be available for validation of currently specified aircraft control resistance values.

Since acceptance is an essential aspect of successful implementation of women into military aviation, the attitudes of a male military population were assessed. The Attitude Toward Women Scale (AWS) devised by Spence and Helmreich (1973) was used to assess general attitudes toward women. The scale was administered to a peer group of 383 male military aviation students and to 47 senior naval officers on the staff of the Chief of Naval Air Training.

As indicated in Table 2, the overall attitudes of the naval officers under instruction (OIs) and aviation officer candidates (AOCs) on the AWS were significantly more liberal than the college normative population, the Marines, and the senior officers. The results suggest acceptance toward women by their potential

Table 2

Selected Statistics for AWS Scores

	College	Marine	Navy		
			AOC	OI	Senior Officers
Mean	89.3	86.4	94.4	99.7	85.2
s.d.	22.5	18.6	19.2	18.4	19.0
N	713	48	211	124	47

peers in military aviation. Acceptance is especially important because of the all volunteer force and continuing need to recruit. The favorable attitudes suggest that the admission of women into military aviation will not have an adverse effect upon male participation. Although the attitudes of the senior officers were more conservative toward women in general, additional data not depicted in the table indicate that they are supportive of plans to extend the rights and opportunities of women in naval aviation.

The ability of an individual to handle multiple complex tasks under stress is an important aspect of military aviation. The Multipurpose Automated Research Station (MARTS) measures such ability. MARTS (Figure 1) is a fully automated test system in which the stimulus elements are mounted behind one-way glass panels and the response keys, except for two floor-mounted controls, are mounted on a desk type console. All stimulus sequences and all responses are presented, monitored, and time-tagged by means of computer.

Three tasks were utilized in this study: a complex-coordination task, a complex discrimination task, and a monitoring task. The complex-coordination task consisted of six stimulus lights and six response switches roughly corresponding to the detection and correction of aircraft roll, pitch, and yaw. Each

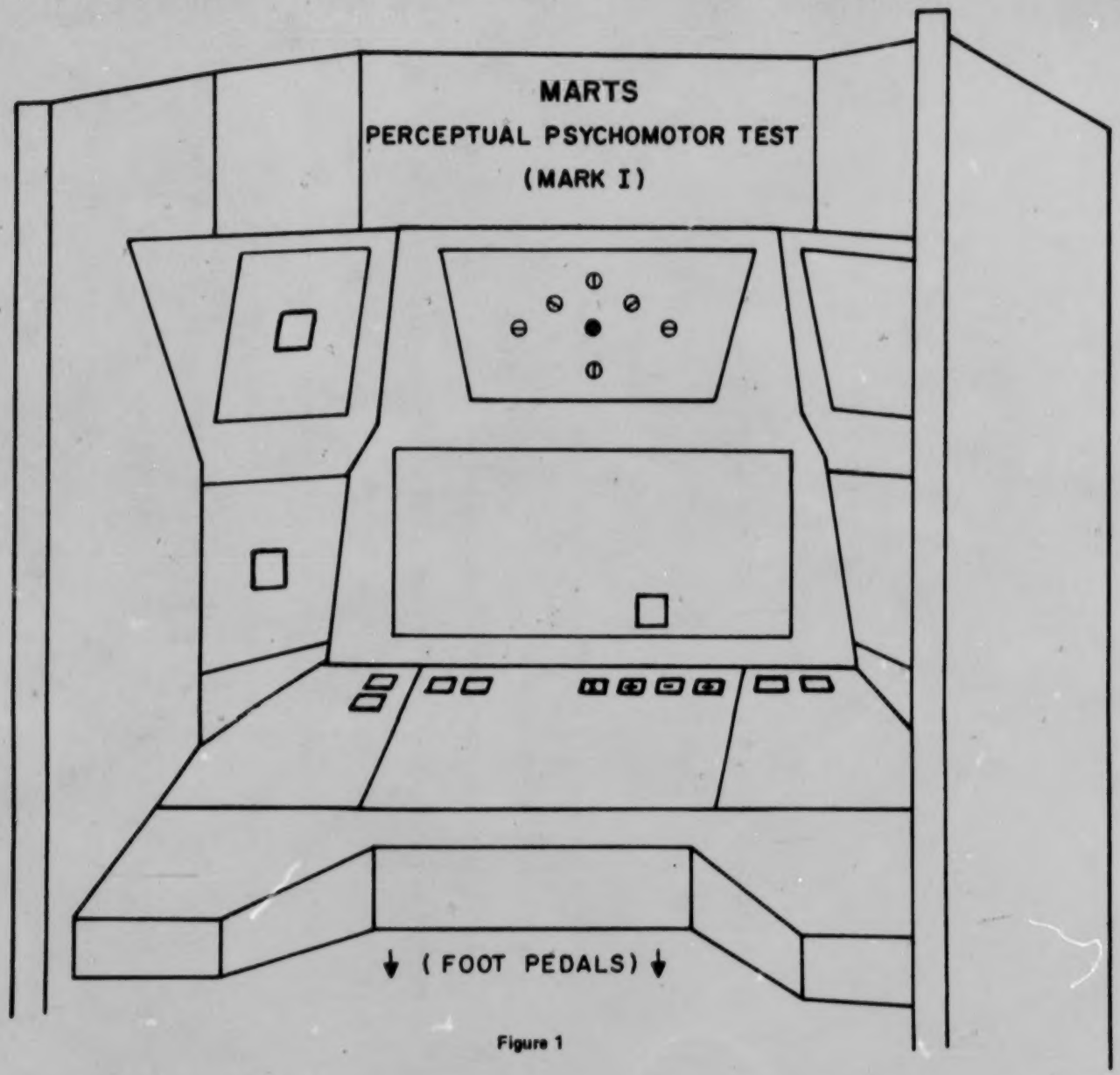


Figure 1

hand controlled two of the response switches and each foot controlled one response switch. Every time a particular stimulus light was illuminated the subject was to turn it off as rapidly as possible by means of the appropriate response switch.

The complex discrimination task consisted of four colored lights paired with four symbols appearing in four colored response keys. The subject was to pair a stimulus color with the response key containing the correct symbol for that color. The symbols appeared on different keys on different trials and, in addition, the colored background of the symbols varied from trial to trial.

The stress test (the warning light monitoring task) consisted of a yellow stimulus light and response switch associated with the indication and avoidance of an impending system malfunction. If the subject made the appropriate response within 4 seconds, the control malfunction, which locked out the response switches to the complex coordination task, could be avoided.

Performance on all tasks was scored in terms of response latencies. The tasks, apparatus, and scoring methodologies are described in greater detail by Gibson (1970).

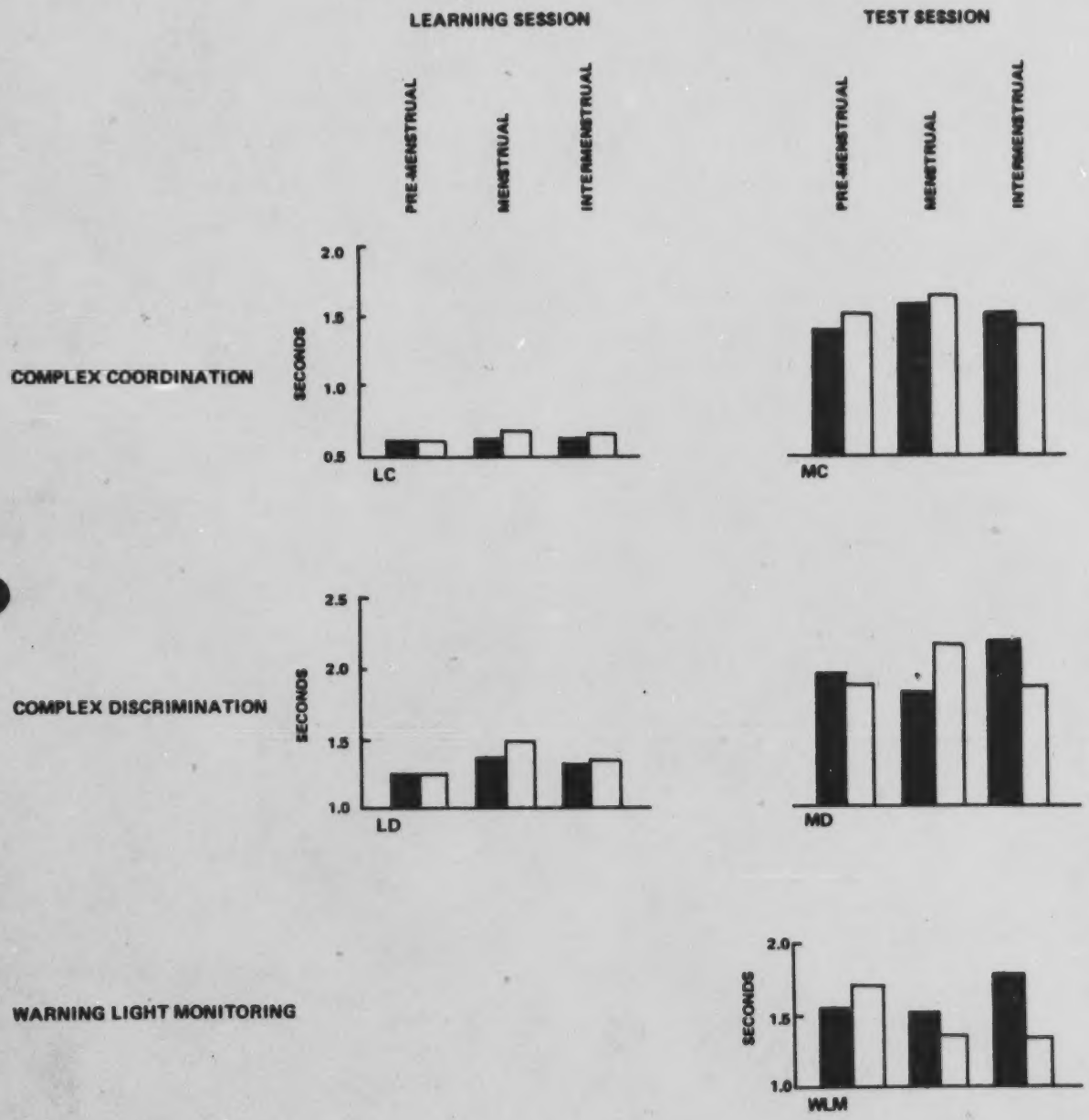
To establish some basis for the female's ability to perform under stress, the performance of the first seven women who entered aviation training was compared to the performance of male flight students. In general, the women performed as well as the male flight students used in the comparison. Intra-subject variability and absolute value of simple decision reaction time were not significantly different for the two groups and were relatively stable under increas-

ing task loading and varying stress event frequency. The percentage of students continuing to respond and the number of responses during increased task loading did not differ. Of the various measures obtained the only significant difference occurred on a subtask requiring a response within 4 sec in order to avoid a control lockout. The females had significantly more lockouts. Unfortunately, the sample size of female aviators is too small to attach any great importance.

Related to the question of stress tolerance limitations for military aircrew women has been the possible effects of the significant physical and hormonal changes associated with the menstrual cycle. This study assesses an individual's ability to perform the previously described complex, perceptual psychomotor tasks during different phases of the menstrual cycle. Seventeen women served as paid volunteer subjects. All were between 20 and 30 years of age with a mean age of 24 years. Each woman was tested during the premenstruum (3 or 4 days prior to menstrual flow), during the menstrual flow, and during the inter-menstrual phase of the cycle. In addition, data were collected on the use of oral contraceptives and on the severity and prevalence of individual menstrual cycle symptomatology as measured by the Menstrual Distress Questionnaire (Moos, 1969). The perceptual-psychomotor data were used to evaluate the effects of menstrual cycle and oral contraceptives on performance under stress. The data were also used to test Dalton's (1964) hypothesis that the lethargy associated with the premenstruum and menstruation is responsible for lowered judgment and increased reaction time. The results of the study (Table 3) indicated that there were no performance decrements associated with the men-

Table 3

MEAN SCORES ON THE COMPLEX PERCEPTUAL - PSYCHOMOTOR TASKS



(557)

strual cycle or use of oral contraceptives and failed to support Dalton's hypothesis (Baisden, 1975).

In conclusion, then, it would seem that the changing scene has imposed considerable demands. It is hoped that studies investigating the aeronautical adaptability of women in military aviation will provide results beneficial to military aviation in general.

A SEGMENTATION ANALYSIS OF WOMEN IN THE ARMY

Lawrence G. Corey
Senior Vice President

Plog Research, Inc.
18631 Sherman Way
Reseda, California 91335

612
551

In a larger study of "Women in the Army" conducted for ARI, two segmentation analyses were conducted using multivariate statistical procedures for defining discrete but homogeneous clusters of men and women, defined by common patterns of attitudes toward women in the Army. The results of these analyses are described under separate subheadings below.

Women's Segmentation Analysis

This analysis proceeded in two steps: first, multidimensional scaling of attitudes toward women in the Army was conducted to develop "root" dimensions on the basis of which a multiple discriminant analysis was performed. The results of these procedures will be discussed under the following subheadings: (a) The Root Dimensions of Segmenting Enlisted Women; (b) Multiple Discriminant Analysis of Women in the Army; (c) The Five Segments of Army Women; and (d) Behavioral Characteristics of Attitude Segments.

...The Root Dimensions of Segmenting Enlisted Women...

Multidimensional scaling of forty-one item battery of attitudes toward women in the Army resulted in producing four root dimensions. The items constituting these roots and their respective loadings are found in Table 1 on the following pages. In summary, these roots can be labeled as (1) Attitudes Toward Combat; (2) Opinions of Other Army Women; (3) Reactions to the Nontraditional MOS; and (4) Competitiveness with Army Men. Each of these is discussed under separate headings below.

Table 1

Root Dimension Loadings for Women's Discriminant Analysis
(N=1718)

		Root I	Root II	Root III	Root IV
1.	A lot of women join the Army because they can't make it in the outside world	.15	.36	.03	.12
2.	Except for the obvious physical differences, a woman and a man are essentially the same	-.36	-.12	-.01	.14
3.	A lot of women in the Army use their sex appeal with male NCO's and officers to get promoted or get special favors and jobs	.08	.23	-.32	.15
4.	When women and men both work on a job, the men should help the women when heavy lifting must be done	.28	.13	.08	.04
5.	In the battlefield, a woman would panic and run under fire	.69	-.00	.04	-.01
6.	Choosing an MOS is strictly an individual thing, and women and men should be free to choose any available MOS their test scores quality them for	-.23	.00	.02	-.03
7.	A woman in the Army is not the kind of woman a man would want to marry	.26	.09	-.05	.16
8.	The moral standards of women in the Army are just as high as those of civilian women	-.38	.25	-.11	.03
9.	A woman isn't capable of fightin in a war	.72	-.04	-.01	-.05
10.	In some ways, women in the Army are sharper and better soldiers than a lot of men I know	-.21	.08	.10	-.07
11.	Most women join the Army because they are looking for a husband	.18	.11	-.07	.20
12.	Men and women in the Army should be treated equally	-.51	-.47	.13	.19
13.	If I had a sister, I wouldn't want her to join the WAC's	.26	.03	.10	-.19
14.	I have a lot of respect for women who go into the "nontraditional" so-called "male" MOS's like mechanic, welder, carpenter, military police, and so forth	-.44	.09	.31	.05
15.	Most of the "nontraditional" MOS's recently opened to women require some heavy lifting which many women can't do well enough	.27	.12	.18	.15
16.	In the Army, men should treat women with tenderness, rather than as equals	.61	.39	.10	.03
17.	Opening "nontraditional," previously "male" MOS's to women wastes a lot of money because most women who try them can't make it and have to be transferred	.46	.15	-.16	.22
18.	Women in the Army get promoted slower than men in the same positions	-.05	-.04	.07	-.01
19.	Most women in the Army are misfits in one way or another	.25	.41	.15	.00
20.	Men and women can fight side by side in the battlefield	-.46	.24	-.07	.30
21.	Women in the Army who get pregnant should be required to get out	.37	.10	-.11	.28
22.	Women should be allowed to enter any MOS a man works in, except combat	.08	-.08	.60	.41

(Table 1 continued)

		I	II	III	IV
23.	The Army is a man's world and women don't really belong in it	.43	.07	-.17	.1
24.	Whether she is in the Army or not, a woman is more fragile and sensitive than a man	.52	.19	.32	.07
25.	A woman in the Army will use her femininity to get away with as much as she can	.10	.25	-.03	.20
26.	NCO's and officers often treat the women they supervise as if they were their daughters	.00	.03	-.04	.07
27.	Women in the Army get promoted faster than men in the same positions	.13	.15	-.14	.20
28.	Women should be careful not to do better than men in competitive situations	.28	-.00	-.03	.15
29.	Most women in the Army won't date men	.13	.19	.03	.16
30.	Women in the "nontraditional" MOS's, such as mechanic, welder, carpenter, military police, and so forth are more masculine than women in the more traditional MOS's	.36	.03	.05	.22
31.	Women should be allowed into combat-related MOS's if they choose	-.61	.12	.15	.05
32.	Women in the Army should be trained in handling weapons	-.57	-.03	..13	.12
33.	Women join the Army for the same reasons men do	-.40	-.05	-.14	.37
34.	Army men treat women in the Army with the same respect as they treat civilian women	-.03	-.03	-.04	.2
35.	Most women in the Army want their NCO's and officers to treat them just like they treat the men	-.32	-.37	-.05	.28
36.	Sometimes an NCO or officer will refuse to promote a woman because she won't date him	-.01	.08	.04	-.15
37.	A woman's monthly period is one of the major reasons she can't be as good a soldier as a man	.43	.13	.19	-.16
38.	Women in the Army draw separate rations because they need a special diet to stay slim	.24	.09	.10	.06
39.	A woman should be kept out of MOS's requiring tough physical exertion because they increase the risk of her getting breast cancer	.44	.07	.25	-.02
40.	Many women in the Army become pregnant just to get a medical discharge from the Army	.15	.01	-.04	.15
41.	It is not safe for a woman to be too successful in the work she chooses to do	.17	.01	.06	.14

Root I: Attitudes Toward Combat

This root is defined by six items, but is most heavily loaded on responses to Items 9 and 5 of the battery:

- o A woman isn't capable of fighting in a war (.72).
- o In the battlefield, a man would not panic and run the way a woman would (.69).

Thus, this root dimension primarily taps the attitudes among women toward their ability to confront and engage in the rigors of armed combat.

Root II: Opinions of Army Women

This root is defined by two items, both of which relate to the extent of negative perceptions of other women in the Army.

- o Most women in the Army are misfits in one way or another (.41).
- o A lot of women join the Army because they can't make it in the outside world (.36).

Consequently, Root II is concerned with perceptions of other Army women as "misfits" and incompetents.

Root III: The Nontraditional MOS

This root is principally defined by three items:

- Women should be allowed to enter any MOS a man works in, except combat (.60).

- A lot of women in the Army use their sex appeal with male NCO's and officers to get promoted or get special favors and jobs (-.32).

Thus, it appears to be measuring attitudes toward traditional and nontraditional MOS's for women and the integrity with which they perform them.

Root IV: Competition with Army Men

This root is defined principally by two items:

- Women join the Army for the same reasons men do (.37).

- Most women in the Army want their NCO's and officers to treat them just like they treat the men (.28).

Thus, it appears that Root IV measures the dimension of competitiveness with men in the Army.

...Multiple Discriminant Analysis...

A multiple discriminant analysis was performed among women, which resulted in identifying five groups or clusters. Table 2 shows the spatial centroids on each root dimension for each of these five groups.

Table 2

Multiple Discriminant Analysis

	Spatial Centroids			
Group	Root I: Combat	Root II: Army Women	Root III: Nontraditional MOS	Root IV: Competitiveness
1	0.82	4.92	3.44	5.66
2	-2.02	2.83	2.38	5.60
3	0.42	2.55	2.88	6.26
4	3.20	2.89	1.87	5.59
5	2.02	1.06	4.35	4.58

These groups and their spatial centroids can be plotted to demonstrate their maximum separation and, thereby, their defining characteristics. Figure A on the following page is an example of one such plot.

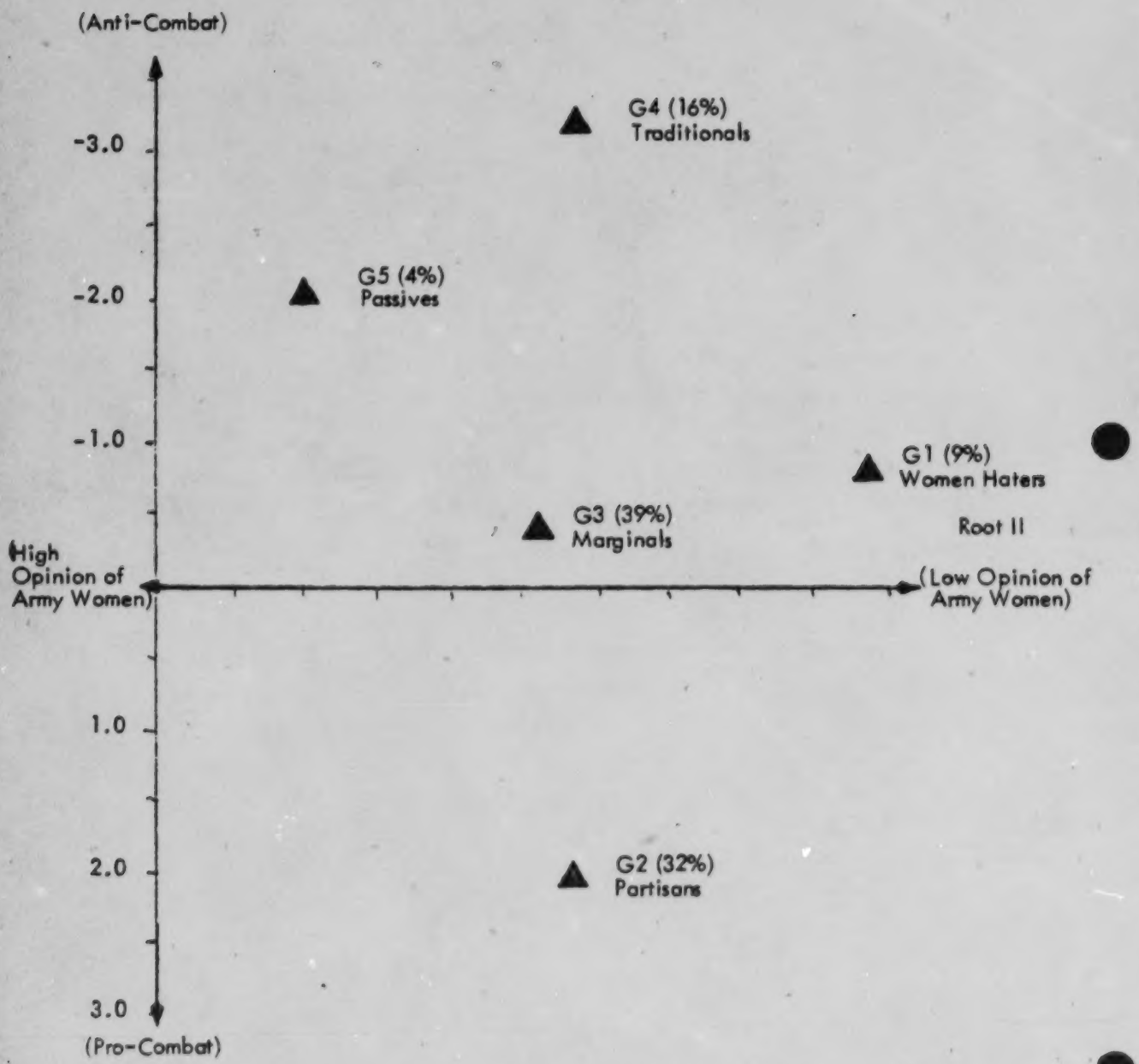
Each group and the attitudes which uniquely define it are discussed separately under individual subheadings to follow.

Group 1: "Women Haters"

This group comprises nine percent of the women in the Army. Referring to Figure A, we find that Group 1 separates from the other four primarily on Root II (opinion of other women) where it evidences a low or negative score. Thus, it appears that Group 1 can be characterized as consisting of enlisted women who have a low opinion of, or are cynical about, other women in the Army.

Figure A

Spatial Centroid Plot of Five Women's Segments



Group 2: "Partisans"

This group comprises 32 percent of all women in the Army. It separates from the other four groups primarily on its affirmative posture toward women's participation in combat.

Group 3: "Marginals"

This group comprises 39 percent of the total female population of the Army. It is defined mainly by its competitive stance toward Army men, but middle position on all other root dimensions.

Group 4: "Traditionals"

This group consists of 16 percent of all enlisted women in the Army. It separates from the other four groups primarily on its stand against combat and the nontraditional MOS for women.

Group 5: "Passives"

This smallest of the five segments accounts for only 4 percent of the female Army population. It separates from the others on its stand against combat, but it has highly favorable attitudes toward other women, accompanied by a lack of competition with Army men.

...The Five Segments of Army Women...

This subsection contains an expanded discussion and description of the five groups of women identified by multiple discriminant analysis. In addition, labels are given to these segments.for purposes of identification. However,

these labels suffer from the usual limitation of factor descriptions and should be taken as only indicative of each segment's general attributes, rather than as definitive characterizations.

Given their separations of their spacial centroids on the root dimensions, as well as other data related to them, the five groups can be labeled as follows:

- Group 1: Women Haters
- Group 2: Partisans
- Group 3: Marginals
- Group 4: Passives

Figure B, on the following page, is a graphic representation of the distribution of these groups in the total population of enlisted women. Each of these groups is described in detail under separate headings.

Group 1: Women Haters

Women in this small segment are characterized by low opinions of other women in the Army. For example, Table 3, on the following pages, shows that 57 percent of the EW's in this segment believe that "a lot of women join the Army because they can't make it in the outside world." This is a significantly higher response than in any of the other segments.

Similarly, women in this segment are far more likely than in any others to agree that "most women in the Army are misfits in one way or the other" (43 percent). Moreover, these Women Haters are much more likely than

Figure B

Segments of Women in the Army
(N=1718)

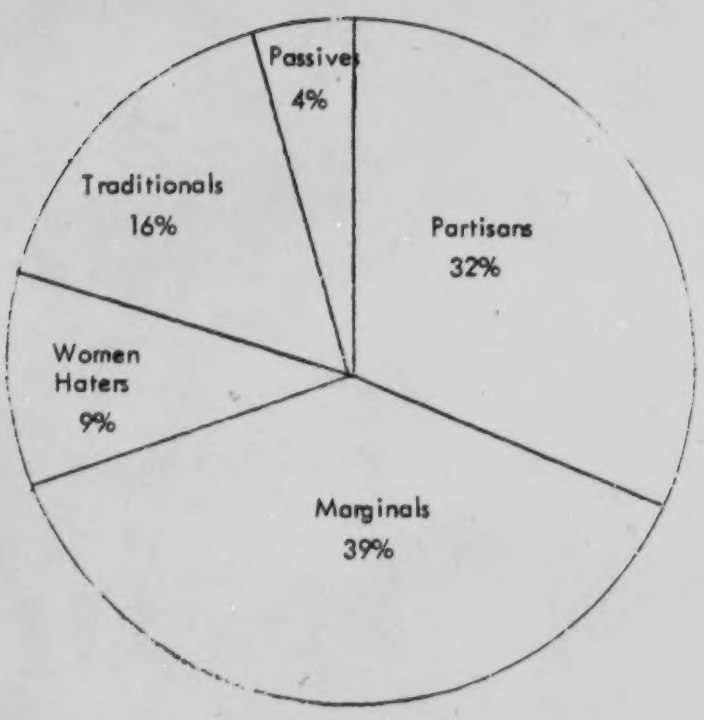


Table 3

Attitudes Toward Women in the Army

	Percent Agreeing					
	Total EW (N=1718)	Women Haters (N=146)	Partisans (N=557)	Marginals (N=670)	Traditionals (N=268)	Passives (N=61)
1. Choosing an MOS is strictly an individual thing, and women and men should be free to choose any available MOS their test scores qualify them for	88.2%	88.4%	93.9%	81.1%	78.0%	80.3%
2. In some ways, women in the Army are sharper and better soldiers than a lot of men I know	82.0	85.6	89.6	81.2	68.7	78.7
3. Women in the Army should be trained in handling weapons	72.3	65.8	92.8	73.6	34.7	59.0
4. The moral standards of women in the Army are just as high as those of civilian women	70.0	79.5	88.3	64.3	48.5	36.0
5. When women and men both work on a job, the men should help the women when heavy lifting must be done	69.2	82.2	54.8	71.2	85.4	78.7
6. Men and women in the Army should be treated equally	61.8	22.0	83.8	69.1	19.5	72.2
7. Women should be allowed into combat-related MOS's if they choose	60.9	63.0	87.2	58.5	18.3	39.4
8. Women join the Army for the same reasons men do	58.6	45.2	74.3	62.0	34.7	24.6
9. I have a lot of respect for women who go into the "non-traditional," so-called "male" MOS's like welder, carpenter, military police, and so forth	55.5	65.8	71.8	54.8	20.1	49.2
10. Whether she is in the Army or not, a woman is more fragile and sensitive than a man	55.0	84.9	26.9	59.7	82.5	80.3
11. A lot of women in the Army use their sex appeal with male NCO's and officers to get promoted or get special favors and jobs	55.0	61.6	55.1	51.6	67.6	19.6
12. Sometimes an NCO or officer will refuse to promote a woman because she won't date him	50.4	58.9	52.5	45.9	53.0	55.7

(Table 3 cont.)

	Total EW (N=1718)	Women Haters (N=146)	Partisans (N=557)	Marginals (N=670)	Traditionals (N=268)	Passives (N=61)
Most of the "nontraditional" MOS's recently opened to women require some heavy lifting which many women can't do well enough	48.0%	68.5%	34.4%	48.7%	61.9%	63.9%
Woman should be allowed to enter any MOS a man works in, except combat	45.2	54.2	33.1	59.7	25.4	73.8
Except for the obvious physical differences, a woman and a man are essentially the same	42.7	28.7	58.0	44.6	17.9	27.8
Most women in the Army want their NCO's and officers to treat them just like they treat the man	39.9	10.3	55.8	42.6	18.6	37.7
If I had a sister, I wouldn't want her to join the WACS	37.5	48.6	27.6	33.8	56.4	62.3
Men and women can fight side by side in the battlefield	35.8	45.2	56.0	30.3	12.7	--
NCO's and officers often treat the women they supervise as if they were their daughters	34.3	34.4	35.9	33.1	33.2	29.5
A woman in the Army will use her femininity to get away with as much as she can	31.8	51.3	28.7	31.8	34.7	16.4
Many women in the Army become pregnant just to get a medical discharge from the Army	30.8	35.7	26.2	31.1	38.8	29.5
Women in the Army get promoted slower than men in the same positions	29.6	29.4	31.6	29.1	26.2	37.7
Women in the "nontraditional" MOS's, such as mechanic, welder, carpenter, military police, and so forth, are more masculine than women in the more traditional MOS's	28.5	35.6	14.0	31.3	46.3	37.7
A lot of women join the Army because they can't make it in the outside world	27.7	56.8	21.8	25.8	34.4	9.8
A woman isn't capable of fighting in a war	20.6	24.0	1.2	15.1	64.2	55.8

(Table 3 cont.)

	Total EW (N=1718)	Women Haters (N=146)	Partisans (N=557)	Marginals (N=670)	Traditionals (N=268)	Passives (N=61)
A woman's monthly period is one of the major reasons she can't be as good a soldier as a man	18.4%	39.1%	5.9%	14.8%	35.8%	44.3%
In the battlefield, a woman would panic and run under fire	18.4	25.3	1.4	13.3	57.5	42.6
Opening "nontraditional", previously "male" MOS's to women wastes a lot of money because most women who try them can't make it and have to be transferred	18.2	28.1	7.1	15.6	45.6	8.2
In the Army, men should treat women with tenderness, rather than as equals	18.0	50.0	2.1	12.3	48.9	14.8
Most women join the Army because they are looking for a husband	13.5	21.2	9.8	11.6	22.4	9.9
Most women in the Army are misfits in one way or another	12.8	43.2	5.6	9.7	19.4	9.9
Army men treat women in the Army with the same respect as they treat civilian women	12.8	15.8	13.7	12.2	11.2	13.1
Women in the Army who get pregnant should be required to get out	11.8	17.1	3.7	11.6	28.3	4.9
Women in the Army get promoted faster than men in the same positions	11.0	20.6	7.4	9.7	17.9	3.2
The Army is a man's world and women can't really belong in it	11.0	14.4	3.6	9.7	28.3	11.5
A woman should be kept out of MOS's requiring tough physical exertion because they increase the risk of her getting breast cancer	10.7	23.3	2.1	8.5	2.4	27.9
A woman in the Army is not the kind of woman a man would want to marry	7.8	12.6	4.3	6.6	16.8	3.2
Women in the Army draw separate rations because they need a special diet to stay slim	7.2	15.7	3.1	6.0	12.3	16.4
It is not safe for a woman to be too successful in the work she chooses to do	6.9	10.9	6.0	6.8	7.5	4.9
Most women in the Army won't date men	6.5	10.3	4.3	5.8	10.1	6.6

(Table 3 cont.)

	Total EW (N=1718)	Women Haters (N=146)	Partisans (N=557)	Marginals (N=670)	Traditionals (N=268)	Passives (N=61)
Women should be careful not to do better than men in competitive situations	4.8%	6.8%	3.2%	4.7%	7.9%	9.8%

other segments to believe that "a woman in the Army will use her femininity to get away with as much as she can" (51 percent). To some extent, they are also more likely to believe that women join the Army "to find a husband" (21 percent) although most women in all five segments disagree with this position.

On the other hand, women in Group 1 are far less likely than those in every group except one to believe that women "want their NCO's and officers to treat them just like they treat the men" (10 percent). This strongly implies they believe most women in the Army want special treatment and perhaps even receive it. Related to this position is the strong disagreement among Group 1 women with the statement that "men and women in the Army should be treated equally" (22 percent). With the exception of one other group (the "Traditionals") this is the lowest level of agreement with that statement.

Thus, the small number of women in Group 1 take a dim view of other women in the Army. They perceive them as primarily opportunistic misfits who attempt to use their femininity to avoid as much as they can.

Group 2: Partisans

This is the second largest segment of women in the Army (32 percent). Women in this group are characterized not only by their highly favorable attitudes toward combat, but also by the degree to which they see women as being tough, capable, and equal to men in almost all respects. For example, Table 3 shows that this group is more likely that any other to endorse the following statements regarding combat:

- Women in the Army should be trained in handling weapons (93 percent).

- Women should be allowed into combat-related MOS's if they choose (87 percent).

- (Least agreement with) In the battlefield a woman would panic and run under fire (1 percent).

- (Least agreement with) A woman isn't capable of fighting in a war (1 percent).

Concerning their opinions about the capabilities of women compared to men, this Partisan segment is more likely than the others to agree that:

- Men and women in the Army should be treated equally (84 percent).

- Women join the Army for the same reasons men do (74 percent).

- Except for obvious physical differences, a man and a woman are essentially the same (58 percent).

- Men and women can fight side by side in the battlefield (56 percent).

- Most women in the Army want their NCO's and officers to treat them just like they treat the men (56 percent).

- (Least agreement with) Whether she is in the Army or not, a woman is more fragile and sensitive than a man (27 percent).

- (Least agreement with) A woman's monthly period is one of the major reasons she can't be as good a soldier as a man (6 percent).

- (Least agreement with) In the Army, men should treat women with tenderness, rather than as equals (2 percent)

- (Least agreement with) The Army is a man's world and women don't really belong in it (4 percent).

By the same token, Partisan women are strongly supportive of other women in the Army. They are much more likely than other groups to agree with the following statements:

- The moral standards of women in the Army are just as high as those of civilian women (88 percent).

- (Least agreement with) When women and men both work on a job, the men should help the women when heavy lifting must be done (55 percent).

And the Partisan is strongly in favor of nontraditional MOS's for women, as indicated by their highest endorsement of the following statements:

- Choosing an MOS is strictly an individual thing, and women and men should be free to choose any available MOS their test scores qualify them for (94 percent).

- I have a lot of respect fo women who go into the "nontraditiona," so-called "male" MOS's like welder, carpenter, military police, and so forth (72 percent).

- (Least agreement with) Most of the "nontraditional" MOS's recently opened to women require some heavy lifting which many women can't do well enough (34 percent).

- (Least agreement with) Women in the "nontraditional" MOS's, such as mechanic, welder, carpenter, military police, and so forth, are more masculine than women in the more traditional MOS's (14 percent).

As later data will show, these positive and affirming attitudes toward other women, combat, and nontraditional MOS's are also reflected in the Partisan's highly favorable attitudes toward the Army. For example, Partisan women are least likely to agree with the statement, "If I had a sister, I wouldn't want her to join the WAC's" (28 percent), thus indicating their approval of Army service for women.

Group 3: Marginals

This is the largest segment of women in the Army (39 percent). They are most characterized by a favorable attitude toward men in the Army, and particularly officers. For example, they are least likely to endorse the Table 3 item, "Sometimes an NCO or officer will refuse to promote a woman because she won't date him" (46 percent).

Interestingly, their spatial centroids fall at the center of each root dimension other than attitudes toward men, which is most favorable. Drs. Jane Prather, Joy Stapp, and Ann Peplau suggest that these relationships indicate Group 3 women are "transitional" or "marginal" in their stand on the issues concerning women in the Army.* Such marginality further implies the possibility of change in one direction or another. Thus, the Marginal segment of women may well represent the best target segment for developing increased participation in nontraditional MOS's.

Group 4: Traditionals

This group represents only 16 percent of the total population of women in the Army. They are characterized by a constellation of attitudes toward women and the Army which are traditional in nature. Generally speaking, they see women as fragile, somewhat subordinate to men, and best suited for traditional MOS's. For example, in Table 3, they are more likely than the other groups to endorse the following statements:

- When women and men both work on a job, the men should help the women when heavy lifting must be done (85 percent).
- A woman isn't capable of fighting in a war (64 percent).
- In the battlefield, a woman would panic and run under fire (58 percent).

* See Appendix H, Volume 4, for the source of this interpretation.

- Opening "nontraditional", previously "male" MOS's to women wastes a lot of money because most women who try them can't make it and have to be transferred (46 percent).

On the other hand, Traditionals are <u>least</u> likely to agree with the following statements:

- In some ways, women in the Army are sharper and better soldiers than a lot of men I know (69 percent).

- Women in the Army should be trained in handling weapons (35 percent).

- Men and women in the Army should be treated equally (20 percent).

- Women should be allowed into combat-related MOS's if they choose (18 percent).

- I have a lot of respect for women who go into the "nontraditional," so-called "male" MOS's like welder, carpenter, military police, and so forth (20 percent).

- Women should be allowed to enter any MOS a man works in, except combat (25 percent).

- Except for obvious physical differences, a woman and a man are essentially the same (18 percent).

Thus, Traditionals not only advocate a separation of men and women, but also the superiority of male troops. They also express some contempt for women who choose a nontraditional MOS.

Group 5: Passives

With only 4 percent, this is the smallest segment of women in the Army. We have labeled this segment the "Passives" because of their rather passive, anti-combat, point of view, coupled with a non-competitive stance toward men. That is, although they do not see women as "misfits," they also do not see them as particularly competitive with men nor capable of facing the challenges of combat.

In terms of their specific attitudes in Table 3, we find that this segment is far more likely than the others to endorse or reject the following statements:

- Women should be allowed to enter any MOS a man works in, except combat (74 percent).
- If I had a sister, I wouldn't want her to join the WACS (62 percent).
- (Absolute rejection) Men and women can fight side by side in the battlefield (no endorsement).
- (Least agreement) A woman in the Army will use her femininity to get away with as much as she can (16 percent).

- (Least agreement) A lot of women join the Army because they can't make it in the outside world (10 percent).

- (Least agreement) The moral standards of women in the Army are just as high as those of civilian women (36 percent).

- (Least agreement) A lot of women in the Army use their sex appeal with male NCO's and officers to get promoted or get special favors and jobs (20 percent).

- (Least agreement) Women join the Army for the same reasons men do (25 percent).

Thus, we see almost a rejection of men and a defensiveness about women among members of Group 5.

The Impact Number of Women in TOandE Units has on Mission Performance

CECIL D. JOHNSON
US Army Research Institute for the Behavioral and Social Sciences

Problem

Women have traditionally been assigned to a limited number of Army positions, medical service and clerical-administrative, for example. In line with other changes in contemporary society, women have agitated for the right and have been allowed to work in less familiar areas in both civilian and military settings. The Army has sought to identify positions that women might fill satisfactorily. Present rules forbid assignment of women to combat type MOS or to TOE units whose wartime mission requires them to operate forward of the brigade rear boundary. Identification of positions in other units has been on an individual basis, with the resulting total of female and interchangeable positions being considerably in excess of what most military experts regard as a realistic proportion of women to allow a combat support or service support unit to perform its mission acceptably. A TRADOC determined limit of 10% for units attached to divisions and 20% for those attached to corps has been proposed, but present fills of TOE combat support units vary from around 0 to 30% women. In order to have "real data" on the effect of women in both service support and combat support units DCSPER has tasked ARI to work with FORSCOM to provide performance measures on units containing varying proportions of women soldiers.

Method of Investigation

Each of 40 TOE companies will be given a TRADOC developed ARTEP (Army Training and Evaluation Program) augmented with ARI developed situations, scenarios and scoring procedures. Fifteen of these companies will be tested twice, once in the fall of 1976 and once in the spring of 1977. Five units will be tested at zero percent women in the fall then filled to 15% for the spring testing. Five others will be tested at 15% in the fall then filled to 35% for the spring testing. A control group of five units will be tested at the same percent of fill with women both spring and fall. The remaining 25 companies will be tested once, sometime in the period between October 1976 and June 1977. The analysis scores will be the mission performance scores obtained from administration of the ARTEP embedded in the test scenario.

WOMEN CONTENT IN UNITS

PURPOSE:

DCSPER NEEDS, BY 1 OCT 77, FIRM DATA ON ACCEPTABLE LEVEL OF FEMALE CONTENT FOR THE ARMY SO AS TO IMPACT, IF NEED BE, ON THE FY 79 FEMALE PERSONNEL LEVEL.

RESULTS MUST:

DEFEND TO THE CONGRESS AND THE COURTS CONSTRAINTS ON SELECTION CRITERIA AND NUMBER OF WOMEN ACCESSIONS.

DEFEND POLICIES AND RULES PERTAINING TO WHERE FEMALES CAN BE ASSIGNED IN PEACETIME AS WELL AS WARTIME.

APPROACH:

THE EFFECT NUMBER OF WOMEN HAS ON UNIT PERFORMANCE MEASURES WILL BE DETERMINED USING TWO TYPES OF COMBAT SUPPORT COMPANIES (SIGNAL AND MP) AND THREE TYPES OF SERVICE SUPPORT COMPANIES (MAINTENANCE, TRANSPORTATION, MEDICAL).

575

WOMEN CONTENT IN UNITS.

SCOPE:

55 STANDARDIZED ADMINISTRATIONS OF ARTEP TO 40 COMPANIES USING SAME 4 PERSON UMPIRE TEAM FOR ALL 8 COMPANIES (11 ARTEP ADMINISTRATIONS) OF THE SAME TYPE. TEST DIRECTORATE WITH ONE COL, 2 LTC's, 5 MAJ's, AND 15 CPT's, WILL COORDINATE AND SCORE MAX WAC ARTEP's. COLLATERAL RESEARCH DATA WILL BE COLLECTED BY ARI STAFF. INSTRUMENTATION, INCLUDING EXPANDED ARTEP SITUATIONS, SCORING PROCEDURES, AND GENERALIZED SCENARIOS ARE BEING DEVELOPED BY ARI --- BASED ON FIVE ARTEP's (ONE FOR EACH COMPANY TYPE) PROVIDED BY TRADOC.

Research Measures

ARTEP

- Overall Mission Performance Measure
  Weighted Sum of Task Scores
- Sub Scores
  Weighted Sum of Selected Task Scores

COLLATERAL RESEARCH INSTRUMENTS

- Background variables (individuals and units)
- Attitudes towards women roles of women
- Differential Standards used by Supervisors
- Confidence in M/F Peers
- Opinions on impact of women on unit effectiveness
- Opinions on validity of ARTEP as measure of effectiveness

Women Content in Units
ARTEP Sub Scores

Task Clusters (Tradition Continuum)	Physical Requirements		Intelligence Requirements	
	High	Low	High	Low
Uniquely Military	X	X		
Other Traditionally Male	X	X		
Traditionally M/F			X	X

Women Content in Units Research Design

Max WAC Units	N IT*	5 Rptd Meas	5 Rptd Meas	5 Rptd Meas	10 Fall Testing	15 Spring Testing
Fill and Training Period		Cross Filling	Cross Filling	No Fills	No Fills	
Fall ARTEP Cycle		Ø % Fill	15% Fill	As Found	As Found	No Testing
Fill and Training Period		Cross Filling & PCS	Cross Filling & PCS	Stabilized Fill		No Fill
Spring ARTEP Cycle		15% Fill	35% Fill	Same as Fall	No Testing	As Found

Content Validity for Training Analysis
(A Practical Application)

by
LCDR C. Lee WALKER
Central Test Site for Personnel and
Training Evaluation Program (PTEP)
Dam Neck, Virginia

In December of 1975 Personnel and Training Evaluation Program (PTEP), which is a Strategic Missile Submarine oriented program, became involved with a training feedback pilot program sponsored by the Chief of Naval Education and Training. In part, the program hoped to measure the effectiveness of a Navy Occupational Task Analysis Program (NOTAP) based training path in supporting fleet needs. The Aviation Electricians Mate (AE) rating which had entry level training ("A" school) based on NOTAP data was selected for evaluation. PTEP was to devise a job knowledge test to measure the AE "A" school graduate 6 months on the job.

PTEP built a test using as a base one hundred twenty seven tasks that AE "A" school had identified as A school training tasks from the NOTAP inventory. Available to support exam construction were the A school test bank and the questions from the Rate Training Courses. The training tasks had first be reduced to underlying knowledge to make a meaningful test of manageable length and more items written. (See enclosure (1) on sample NOTAP tasks). Two paper work fault isolation exercises were also included in the test plan. A submarine Electronics Technician had the principle responsibility for test design and construction working with a great deal of support from local Aviation Electrician's Mates.

Although this methodology for test development was in keeping with academic test writing practices it was determined to use a Content Validity measure on the test on a trial basis. The method chosen was one presented by DR C. H. LAWSHE of Purdue University at a seminar at Bowling State University in July of 1975. The method had each of several subject matter experts determine for each item:

Is the skill (or knowledge) measured for this item

--Essential
--Useful, but not essential
--Not necessary
To the performance of the job.

A content validity ratio is then computed using the following formula

$$CVR = \frac{n_e - N/2}{N/2}$$

in which n_e is the number indicating "essential" and N is the number of panelists.

The first attempt at content validation consisted of placing each item on a single card grouped by task supported. Panelists were then asked to place the cards in one of three envelopes. No request was made for reason for assignment. It was hoped that this would be a quick and simple procedure. The Senior Chief Electronics Technician returned from his attempt at content validation with some help on specific items but no general validation. The test was then piloted at replacement and advanced training sites. The results of the pilot were disturbing. AE personnel were having considerable difficulty with a test which the builder thought was easy in the extreme. Either the test was on irrelevant items or was valid and identified a training problem. Prior to general administration therefore content validity had become a real issue.

It was felt essential that the content validity of the test be documented so that the diagnostic results could be forcefully presented. A renewed effort was made to carry out DR LAWSHE's Content Validity procedures. Ten enlisted subject matter experts were used plus two retired Aviation Electrician Mates working in the training field. Each man was given a listing of questions with the specific task to which they were related. He was asked to mark a multiple choice answer sheet 1, 2, or 3 for each question and to provide comments where he wanted. Background data was gathered on each panelist. This approach proved more acceptable then the card sort and the data was returned with little difficulty. DR LAWSHE's formula and criterion were applied to the results of the poll and 76 of 100 items cleared. That left 24 questions and some six weeks of completed work for further evaluation. Response patterns for individual panelists were examined and arbitrarily weighted. (Enclosure 2). Two panelists it proved were providing the bulk of the objections, other questions proved to be aircraft type dependent. Based on the second sort we eliminated some questions and retained others. The resulting examination is felt to be one well suited to its purposes with a sufficient recorded history of validation that it can be used as a mover for supporting rather significant changes in A school curricula.

Between the first and second attempts to apply DR LAWSHE's content validity criterion Central Test Site received a procedure for selection device content validation which had been prepared by DR John MENNE of Iowa State University. This device provides a convenient and practical way for recording and presenting the content validation procedure starting with a validation of tasks and proceeding through to individual items used as a basis for selection. A four point rating scale for items is used in this device and no formula or mathematical criteria is applied for item acceptance (Enclosure 3).

Training evaluation methodology being investigated by the Navy in support of a system approach to training includes several approaches using measurement instruments. For credibility of the resulting measurements with technical personnel at all levels it is important that the Content Validity of the measurement instrument be clearly established. DR LAWSHE and DR MENNE's work both provede a practical method for establishing and documenting Content Validity.

Duty Title	RECOMMENDED TRAINING ENVIRONMENT		
	P	A	C
M11 Test/check accelerometers		X	X
M14 Test/check flight control trim system		X	X
M16 Test/check flight control system		X	X
M18 Test/check wing anti-ice system		X	X
M19 Test/check windshield de-ice system		X	X
M25 Test/check turn and bank indicators		X	X
M26 Test/check true airspeed indicating system		X	X
M27 Test/check attitude indicating system		X	X
M28 Test/check airspeed indicators		X	X
M29 Test/check a/c quantity indicating systems		X	X

Enclosure (1)

(588)

	AEC INST	AEC A-6 INST	AEC INST	AE1 INST	AEC	AE1	AE2 F-14	AE1 F-14	AEC F-14	AE2 F-14
	4/1	7/7	7/4	0/0	12/1	20/9	34/20	8/5	2/5	3/8
11						X	1	1		
12						1	1		X	
15		1	1				X	1	1	
16						1	X			X
22					1	1				
25	1		X			X	X	1		X
28	X		X			1	X	1		
31		1	1				1			
32	1	1				1	X			
33	1		1				X			
35						1	X	X	X	X
36							1	1		X
37							1	1		
46		1			1	1	X			
55						X	1	1	X	
60		1	1			1	1		X	X
62		X				1	1			
64		X				1				
67					1	1	1			
68					1	1	1			
89		X				1	1			
92						X	X			1
96		1				1	X			
97	1		X		1	1	X			X

ITEM RATING SCALE

1. UNRELATED: NOT CLOSELY RELATED TO ANY KNOWLEDGE, ABILITY, SKILL, OR PERSONAL CHARACTERISTIC NEEDED FOR THIS JOB.

2. FAIR: related to the level of ____ (record the appropriate KASPC Reference Code on the Item Rating Form) needed for adequate job performance, but not likely to make significant distinctions between adequate and inadequate levels of competence.

3. GOOD: related to the level of ____ (record the appropriate KASPC Reference Code on the Item Rating Form) needed for adequate job performance, and likely to make significant distinctions between adequate and inadequate levels of competence.

4. SUPERIOR: related to the level of ____ (record the appropriate KASPC Reference Code on the Item Rating Form) needed for adequate job performance, and likely to make significant distinctions between adequate and superior levels of competence.

NOTE: DO NOT RECORD A KASPC REFERENCE CODE FOR ITEMS THAT YOU RATE "1".

Enclosure (3)

JOB ANALYSIS QUESTIONNAIRE FOR

SELECTION DEVICE CONTENT VALIDATION

Iowa Merit Employment Department
Grimes State Office Building
Des Moines, Iowa 50319

DR John W. MENNE
Test and Evaluation Services
44 Carver Hall
Anes, Iowa 50010
515-294-2232

DR C. H. LAWSHE
Personnel Psychology
1975-28-563-575

THE DEVELOPMENT OF STATISTICALLY PARALLEL TESTS BY ANALYSIS OF UNIQUE VARIANCE

Malcolm James Ree

AIR FORCE HUMAN RESOURCES LABORATORY
PERSONNEL RESEARCH DIVISION
LACKLAND AIR FORCE BASE, TEXAS 78236

October 1976

584 646

I. THE PROBLEM

Most problems of testing concern the validity or the reliability of measurement. Validity is usually determined by the correlation coefficient although other methods are used from time to time. Reliability, on the other hand, is estimated in a variety of ways, for example, split half, alternate forms, or test-retest. The parallel forms reliability coefficient has many advantages. It is the reliability coefficient which accurately shows the proportion of true score variance to total variance. Few assumptions are required; only that the test forms measure the same factor or factors. The parallel forms coefficient is a Pearson product moment correlation and may be tested as any other Pearson product moment correlation. Further, the standard error of measurement may be computed directly from the distribution of difference scores based on the two forms.

Parallel forms reliability is often not computed because it requires at least two forms which have equal means, equal variance, and equal correlation with a criterion. Logically, the items in both forms must sample the same universe; thus a mathematics test and a spelling test, regardless of equal means, variance, and criterion correlation, cannot be parallel forms.

The difficulty usually associated with estimating reliability by parallel forms is building the forms from pools of items. Analysis of unique item variances is a procedure which allows parallel forms to be established from a limited number of items. The procedure for this was developed in order to produce two parallel forms from 120 right-wrong scored perceptual items. This new procedure was developed because traditional methods of producing parallel forms, such as random assignment of items, had failed several times to yield rigorously parallel forms from these items.

II. METHOD

Subjects

A group of 907 basic airmen trainees at Lackland AFB, Texas was selected on a random basis. The pool of 120 items was administered to the subjects in one continuous time period.

These subjects were randomly assigned to a developmental group of 350, a reliability group of 350, and a validity group of 207. Therefore, in each case, cross-validated test forms were tested.

Items

The items were very similar in form. Each required the subject to estimate the angle at which an object in a photograph was viewed. The subject was asked to select the proper angle from among eight angles presented. The same set of eight angles from 0° to 90° served as the answers for each item. Past experience indicated that these angle estimation items formed a homogeneous group (Davis, 1957).

Criterion Measure

The validity criterion was a computer driven and scored free-angle estimation task. The subject was required to push buttons on a keyboard which corresponded to his estimate of the angle-of-view of a series of line-drawn figures. Scores on this task were the algebraic sum of errors for each of ten items. These scores were subtracted from a constant for ease of interpretation and then converted to a unit normal Z score.

Procedure

Ranking Items by Uniqueness. A FORTRAN program known as VARSEL was used to rank all the items by unique variance. VARSEL is described in detail elsewhere (Gould & Christal, 1975), but a brief description is offered here.

VARSEL first computes all the R^2 of each item versus all the remaining items; for example, 1 versus 2 through 120, 2 versus 1 and 3 through 120, 3 versus 1, 2, 4 through 120, etc.

Then the unique or unaccounted-for-variance is computed by

$$U^2 = 1 - R^2 \quad (1)$$

where U^2 is the unique variance.

The item with the highest unique variance is put into a pool *P*. This item is then correlated against each of the remaining items. From among these resulting U^2's, the item with the highest U^2 value is added to the pool *P*.

Next, the best weighted combination of items in the pool is correlated with each of the remaining items, and the highest value of U^2 determined. The item which yields the highest U^2 then becomes the next member of *P*. This procedure is repeated until all of the items have been included in *P*; however, the procedure can be made to stop on any iterative step.

The items are now ranked for uniqueness by order of inclusion from most unique variance to lowest unique variance.

Assigning Items to Tests. Logically, test forms which are statistically parallel should have a high common variance between the scores. This is usually measured by the squared correlation of the scores of the two parallel forms. The task is to allocate items to forms so that no form has too much unique variance.

To allocate the items, it was reasoned that assigning the items selected for the pool on odd-numbered iterations to one form and the items selected for the pool on even-numbered iterations would apportion the unique and common variance equally. The two forms were created in this manner and scored for the reliability and validity groups by the sum of the rights-only method.

Statistical Approach

In order to investigate the effects of the procedure on reliability and validity, the following tests were made:

Reliability. Parallel forms reliability was estimated by computing the correlation between the raw scores on the two forms. This was tested by the following hypothesis:

a. HO: $R_{AA'} = 0$

H1: $R_{AA'} > 0$

The means and variances of the two forms were tested to determine if they differed between forms by the following hypotheses:

b. HO: $\bar{X}_A = \bar{X}_{A'}$

H1: $\bar{X}_A \neq \bar{X}_{A'}$ and

c. HO: $V_A = V_{A'}$

H1: $V_A \neq V_{A'}$

Hypotheses a and b were tested by forming a t-ratio, and hypothesis c was tested by forming an F-ratio.

Validity. The scores on the two forms for the members of the validity group (N = 207) were correlated with the scores on the criterion. Several hypotheses were of interest. These were

d. HO: $R_{C.A} = 0$

H1: $R_{C.A} > 0$

where C is the criterion score and A is test form A,

e. HO: $R_{C.A'} = 0$

H1: $R_{C.A'} > 0$

where A' is test form A' and finally,

f. HO: $R_{C.A} = R_{C.A'}$

H1: $R_{C.A} \neq R_{C.A'}$

These hypotheses were each tested by forming a t-ratio. It can immediately be seen that failure to reject the null hypothesis from b, c, and f is the usual test to determine if two forms are parallel. Hypothesis a, d, and e were tested prior to b, c, and f to preserve logic. These five hypotheses were considered as a group, and Type I error was taken for the entire group at $P < .01$.

III. RESULTS

Reliability

Table 1 presents the means, variances, and other descriptive statistics for the two parallel forms for the reliability group.

Table 1. **Descriptive Statistics for Forms A and A' for the Reliability Group (N = 350)**

Descriptors	Form A	Form A'
Number of items	60	60
Mean	24.71	25.20
Standard Error of the Mean	.44	.46
Standard Deviation	8.15	8.58
Variance	66.44	73.57
Coefficient Alpha	.81	.83

The t-ratio which was constructed to test the difference between the means was .893 (df = 348). Thus, the null hypothesis could not be rejected.

The F-ratio which was constructed to test the equality of the variances was 1.107 (df = 348). The null hypothesis was again not rejected.

The parallel forms reliability coefficient of .825 was tested to determine if it was significantly greater than zero. The t-ratio formed permitted the null hypothesis to be rejected.

Validity

In order for the two forms to be considered parallel, each had to correlate equally with a criterion. Form 1 was found to correlate with the criterion at .2426, and form 2 was found to correlate with the criterion at .2418.

Each was tested to determine if it was significantly greater than zero. The t-ratios for form 1 and form 2 were, respectively, 3.579 and 3.517 (df = 205). Both correlations were found to be significantly greater than zero.

A t-ratio was also computed to determine if these two correlations differed from each other (McNemar, 1949, p. 125). The obtained value was 1.513 (df = 204); thus, the hypothesis could not be rejected.

IV. DISCUSSION

The tests of the hypotheses indicated that the two forms of the Angle Estimation Test were indeed parallel. The means were equal, the variances were equal, and there was equal correlation with a criterion.

The finding of statistical parallelism is insufficient for determining test forms truly parallel. The one assumption in parallel forms estimation of reliability is that the tests measure the same attribute or

combination of attributes. It is surely illogical to impute parallel forms status to two tests that measure different factors, no matter how statistically interchangeable the scores may be. Kelley (1942) presents a fine discussion on this subject.

It was not difficult to make the assumption that all the items in the group of 120 were measures of the same attribute. The items were all of the same type and form, specifically, pictures of models. The task was the same for each item and the possible answer set was the same eight angles. The test constructors (Davis, 1957; Fruchter, 1975) specified that these items were created to measure a single ability.

It was reasoned that if two parallel forms could be produced by an odd-even split, then further splitting of the scales could produce additional shorter parallel forms. Eight 15-item forms were produced in this manner. Table 2 presents descriptive statistics for the eight 15-item scales. Neither the means nor the variances differ between forms.

Table 2. **Descriptive Statistics of the Eight Scales**

Scale	Mean	Standard Error of the Mean	Variance
1	5.82	.18	6.61
2	5.28	.17	5.80
3	5.84	.17	6.23
4	5.86	.18	7.09
5	6.08	.18	6.79
6	6.07	.18	7.12
7	6.28	.19	7.27
8	5.69	.18	6.79

A factor analysis of the intercorrelation was carried out using the method of Principal components and the Varimax rotation. Only one Eigen value of greater than 1.0 was found. The one factor accounted for 64.8 percent of the variance. Table 3 shows the rotated factor loadings of subtests.

Table 3. **Rotated Factor Loadings of the Subtests**

Variable	Factor I
1	.84
2	.77
3	.79
4	.85
5	.79
6	.80
7	.81
8	.79

The loadings for each scale were all very similar. This indicated that each of the scales was measuring the same factor. This was further evidence that the VARSEL procedure could be used to allocate variance to scales in a manner that produced statistically parallel test forms.

In order to demonstrate that the technique's success was not a function of the nature of the angle estimation items, a replication on a right-wrong scored attitude measure was done. A supervisor's rating served as the validity criterion. Table 4 presents the results of this replication.

Table 4. **Results of Replication on Scored Attitude Scale (N = 259)**

Descriptors	Form A	Form B
Number of Items	23	23
Mean	8.58	7.53
Variance	16.82	16.77
Validity	.21	.26

Again, the five hypotheses were tested with a group Type I error rate of $P < .01$. The tests were found to be statistically parallel.

The results of the replication added evidence to the utility of the procedure.

V. CONCLUSIONS

The procedure of producing statistically parallel test forms by analysis of unique item variance is successful. When the need arises for parallel forms, they can be produced from pools of items which are logically homogeneous.

REFERENCES

Davis, F.B. *Final report–Air Force Contract AF 41(657)-66,* 1957. Bronxville, N.Y.

Fruchter, B. Personal correspondence, 1975.

Gould, R.B., & Christal, R.E. *VARSEL: Variable selection for multiple-purpose prediction systems without using external criteria.* Paper presented at the annual meeting of the Military Testing Association, Indianapolis, IN, September, 1975.

Kelley, T. The reliability coefficient. *Psychometrica,* 1942, 7, 75–83.

McNemar, Q. *Psychological statistics.* New York: John Wiley and Sons, Inc., 1949.

THE SEARCH FOR FAIRNESS IN TESTING

by
Raymond L. Lewiski
Education Specialist

Extension Course Institute
Air University

ABSTRACT

This article is divided into two parts: the historical prelude and the search for fairness. The article begins with a brief historical acknowledgement of noteworthy events that have fostered a general sense of fairness within our cultural heritage. Specifically cited are the 1964 Civil Rights Act and the Equal Employment Opportunity Act of 1972. These are placed in perspective with recent events that have impacted on the notion of fairness as it relates to employment testing. The article shows how these legislative acts set the ground rules that are to be followed in the search for fairness in testing. The historical prelude is concluded with fairness in testing viewed as a legal mandate as well as a desirable social goal.

Next, the search for fairness in testing is begun by a review of noteworthy axiological models of fairness in testing that have been proposed during the last ten years. For each model reviewed, the concept of fairness that logically follows is described and the dominant value commitment associated with the model is identified. Finally, each model is applied to a hypothetical employment situation and its dominant value is highlighted.

THE SEARCH FOR FAIRNESS IN TESTING

Historical Prelude

The notion of fairness as a social goal is steeped in American tradition. The Magna Carta, the Declaration of Indpendence, the Emancipation Proclamation, the Civil Rights Act of 1964, and the Equal Employment Opportunity Act of 1972 are all social events that have advocated a sense of fairness and shaped the American heritage. Although each of these events has contributed to a general sense of fairness, the issues of fairness in testing were not addressed until 1964.

The 1964 Civil Rights Act and the Equal Employment Opportunity Act of 1972 are of particular importance to fairness in testing as sections of both acts have dealt directly with the question of how tests are to be used. During 1964, Civil Rights legislation was being debated in Congress and the question of unlawful employment practices was also raised and debated. During debate on this issue, an amendment was introduced by Senator Tower of Texas (later to become the Tower Amendment to Title VII of the Civil Rights Act of 1964). The Tower Amendment was embodied in Section 703(h) of Title VII of the act and reads as follows:

> Nor shall it be an unlawful employment practice for an employer to give and act upon the result of any professionally developed ability test provided that such test, its administration or action upon the results is not designed, intended or used to discriminate because of race, color, religion, sex, or national origin.[1]

The Tower Amendment has had considerable impact on public policy toward the use of tests, and has been the center of much controversy. Shortly after the 1964 Act became law, the wording of the amendment led to a controversy over the use and meaning of the term "ability test." The question was raised of whether a legal interpretation of the term "ability test" was to include broad general abilities that tend to measure the person in the abstract, or be restricted to specific job-related abilities that could be empirically verified. In other words, were mental ability tests or intelligence tests also classed as "ability tests"? Members from both the psychological and legal professions were in disagreement over how the term "ability test" was to be interpreted. An answer to this question was not easily forthcoming.

The controversy was settled in part by the case of Willie S. Griggs et al., Petitioners., vs Duke Power Company. The Griggs vs Duke Power Company case was decided in the U. S. Supreme Court on 8 March 1971. The court ruled that tests used for the purpose of making employment decisions were required by law to be job-related, and the burden of proving job-relatedness was placed on the employer.

Justice Burger interpreted Congressional intent as follows:

> Nothing in the act precludes the use of testing or measuring procedures; obviously they are useful. What Congress has forbidden is giving these devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance. Congress has not recommended that the less qualified be preferred over the better qualified simply because of minority origins. Far from disparaging job qualifications as such, Congress has made such qualifications the controlling factor so that race, religion, nationality, and sex become irrelevant. What Congress has commanded is that any tests used must measure the person for the job and not the person in the abstract.[2]

This historic decision provided a basis for one of the ground rules of employment testing: a test must be job-related. Other ground rules resulted from guidelines that were written by agencies established by the 1964 Civil Rights Act and the 1972 Equal Employment Opportunity Act.

The Equal Employment Opportunity Act of 1972 (Public Law 92-261) amended Title VII of the Civil Rights Act of 1964, expanding its coverage and adding enforcement powers. Section 706(a), as amended, empowered the commission to prevent any person from engaging in any unlawful employment practices as set forth in Section 703 (Discrimination because of Race, Color, Religion, Sex, or National Origin) and Section 704 (Other Unlawful Employment Practices). Another change affected the Attorney General's Office. Cases involving a government, governmental agency, or political subdivision were charged to the Attorney General. Section 707(c), as amended, provided for the commission's assumption of the functions that were previously charged to the Attorney General and his office. These functions were to be assumed within two years after the date of enactment of the Act.

Another important portion of the 1972 EEO Act was Section 715. This section provided for the establishment of the Equal Employment Opportunity Coordination Council (EEOCC) whose membership is composed of the following five heads of government agencies or their respective delegates:

1. Secretary of Labor
2. The Chairman of the Equal Employment Opportunity Commission.
3. The Attorney General
4. The Chairman of the United States Civil Service Commission.
5. The Chairman of the United States Civil Rights Commission.

The responsibilities of the EEOCC are described in Section 715 as follows:

> The Council shall have the responsibility for developing and implementing agreements, policies and practices designed to maximize effort, promote efficiency, and eliminate conflict, competition, duplication and inconsistency among the operations, functions and jurisdictions of the various departments, agencies and branches of the Federal government responsible for the implementation and enforcement of equal employment opportunity legislation, orders, and policies. On or before July 1 of each year, the Council shall transmit to the President and the Congress a report of its activities, together with such recommendations for legislative or administrative changes as it concludes are desirable to further promote the purposes of the section.[3]

The extent to which the EEOCC has been able to carry out its responsibility is an open question. A new set of guidelines has been proposed and was published in the Federal Register on the 14th of July 1976. However, no attempt was made to obtain concurrence from one member, the Civil Rights Commission, on the proposed guidelines, and another member, the Equal Employment Opportunity Commission, maintains that the guidelines do not represent the position of that agency. In spite of their differences, all members of the EEOCC agree that a uniform set of guidelines ought to be established.

In the final analysis, both the 1964 Civil Rights Act and the 1972 Equal Employment Opportunity Act have had considerable impact on the way tests are to be used, and each has contributed to the establishment and maintenance of a sense of fairness with respect to testing. In effect, these acts have placed the issue of fairness boldly before the test user. A contemporary user of tests may view these acts as desirable and needed social goals or as a legal obligation. Regardless of point of view the user cannot escape the issue of fairness.

The Search for Fairness

At first encounter, the notion of fairness is a nebulous concept that permeates both the employment decision-making process and the possible consequences that may result from the decisions that are made. The question of fairness may be raised at any stage of what may be referred to as the overall selection process;however, some points within the process are more crucial to fairness than others.

The Overall Selection Process

An outline of the overall selection process is given in figure 1 where the selection process is divided into two stages: a decision-making stage and a possible-outcomes stage. The decision-making stage is subdivided into three substages comprising tasks to be performed. These are as follows: (1) the identification and stratification of the population for which an employment decision is to be made. The Uniform Guidelines on Employee Selection procedure suggest that the population be identified and stratified by sex, race, and ethnic groupings. Also, the Guidelines define employment decisions as follows:

> Employment decisions include but are not limited to hire, additional compensation or promotion, demotion, membership (for example in a labor organization), referral, retention, licensing and certification, to the extent that licensing and certification may be covered by Federal equal employment opportunity law. Employment decisions may also include transfer, job or work assignments, or training if they are used as a basis for additional compensation.[4]

(2) the gathering and validation of types of information to be used for making employment decisions. According to the Uniform Guidelines on Employee Selection procedure, this includes the following types of information:

> Any measure, combination of measures, or procedure, other than a bona fide seniority system, used as a basis for any employment decision. Selection procedures include the full range of assessment techniques from traditional paper and pencil tests, performance tests, training programs or probationary periods and physical. Educational and work experience requirements through informal or casual interviews and unscored application forms.[5]

(3) the identification and use of strategies for decision making. At this substage, each of the axiological models of fairness reviewed in this paper is an example of a unique strategy.

Insofar as possible, the tasks to be performed in the decision-making stage are logically ordered. For instance, the requirement that there be a population on which an employment decision is to be made is logically prior to the task of gathering information. Likewise, before a decision-making strategy can be used, there must be valid information available. Also, it may be noted that the logical ordering of the tasks would permit fairness to be jeopardized before its consideration at a later stage of the overall selection process. For example, if the prior task of test validation were not properly performed, the later task of making an employment decision has dubious fairness.

The second stage within the overall selection process consists of a breakdown of possible outcomes. These are shown in the lower part of figure 1. Possible outcomes impact on either an individual or a job to be performed, and an analysis of outcomes is considered essential to any notion of fairness. The first possible outcome is that a person is initially selected for favorable treatment or not selected. Favorable treatment is considered as any change in a person's situation or status that is advantageous for that person. Being selected for favorable treatment does not mean that one will reap the benefits of a change in his or her situation. For example, with respect to job openings, there may be more job openings than there are qualified candidates to fill them or there may be more qualified candidates looking for jobs than there are job openings. In the former case, all persons selected for favorable treatment would most likely reap the benefits of an improved situation, while in the latter case, some would not. The final outcome considered with respect to recipients of favorable treatment pertains to the beneficiary's later performance on the job. Here there are two possibilities: a person's job performance will be either acceptable or not acceptable. For those persons whose situations remain the same, their final outcome may be reviewed from various points of view: psychological, social, economic, legal, political, etc.

THE OVERALL SELECTION PROCESS

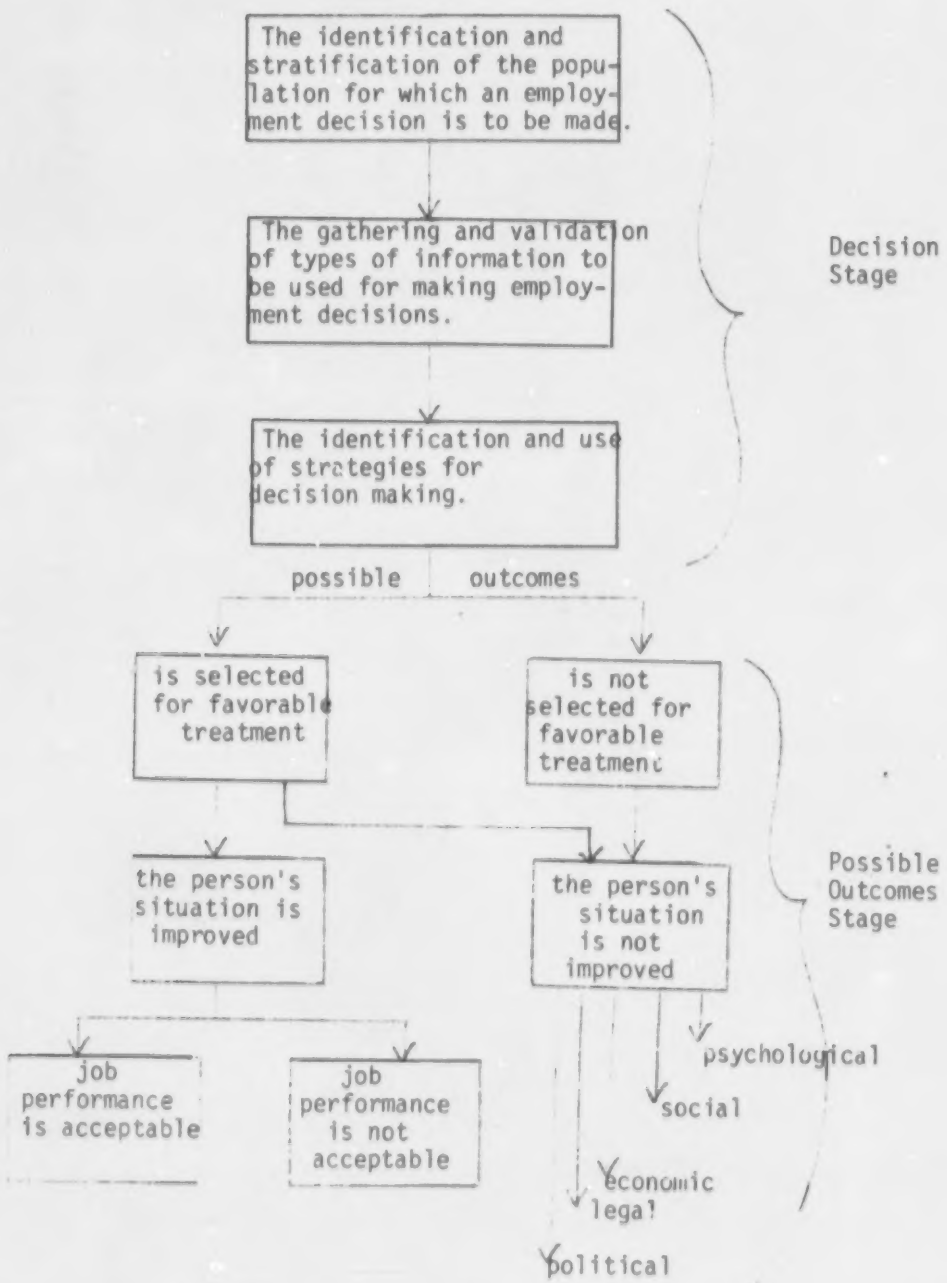

The identification and stratification of the population for which an employment decision is to be made.
The gathering and validation of types of information to be used for making employment decisions.
The identification and use of strategies for decision making.
Decision Stage
possible outcomes
is selected for favorable treatment
is not selected for favorable treatment
the person's situation is improved
the person's situation is not improved
Possible Outcomes Stage
job performance is acceptable
job performance is not acceptable
psychological
social
economic
legal
political

Figure 1

A thorough search for fairness would require close scrutiny of all stages and substages of the overall selection process. This paper will consider only that portion of the selection process that begins at substage 3 of the decision stage and continues through most of the outcome stage. In all cases in which fairness is discussed, it will be assumed that the first two tasks in the decision stage have been properly performed. In a real life situation, this assumption would not be permitted but for the hypothetical situations that will be considered here the assumption is in order.

Currently, the most discussed and debated segment of the overall selection process is the third substage of the decision-making stage. It is at this substage that most disagreement begins. Most professional developers and users of tests agree with the social goal of fairness in testing and welcome the legal requirements of an affirmative action program. They disagree, however, on what constitutes "fairness in testing" and how "fairness" can best be measured.

During the last ten years, a number of definitions have been suggested for "fairness in testing" and various models have been proposed for measuring fairness best. Seven such models will be reviewed in this paper. This review will be made from a practitioner's point of view. No attempt will be made to spawn a new model of fairness. Also, insofar as possible, the review will be literary and at all times an attempt will be made to minimize the use of mathematical symbols. An attempt will be made to answer questions about past models used, what is being done now and what we can expect for the future.

For each model reviewed, the dominant value commitment associated with the model will be identified and the concept of fairness that logically follows will be described. Then, each model will be applied to a hypothetical employment situation and its dominant value will be highlighted.

The Traditional Selection Strategy

Before the 1964 Civil Rights Act was passed, relatively little had been written about fairness in testing and the issue of the fairness of a selection strategy had not been raised. At this time, it was generally concluded that a selection strategy based on a simple regression analysis was essentially fair and the only axiological model seriously advocated was a straightforward, uncomplicated regression model. Later variations of this model were to consider the applicant's sex, race, and ethnic affiliation and were more complex. The traditional model was based on a simple regression analysis in which the regression of a job

performance criterion Y on a text score X was determined for a general population first; then, after having established the regression of Y on X, the resulting regression weights were used to predict an expected job performance score Y' for each new job applicant. Letting (a) equal the intercepts and (b) equal the slope, the simple regression equation was $Y' = a+bX$.

The regression model has traditionally been advocated by people who were primarily concerned with the maximization of their ability to select job applicants who would most likely be successful on the job. Consequently, this model favors the high scoring applicant on text X. Using this strategy, the applicant with the highest predicted score Y' would be hired first, the next highest second, and so on till all job vacancies were filled. This is an ideal selection strategy from an employer's point of view but, for applicants concerned with their equal employment opportunity or a group-oriented individual, a different strategy would most likely be preferred.

After the 1964 Civil Rights Act became law, the basic questions of fairness in testing and selection were quickly raised in court, and it was during this period time that fairness of the regression model was first seriously challenged. Three criticisms were leveled against traditional selection strategies. It was argued that they were inappropriate for dealing with social conditions created by a long past history of job discrimination. It was repeatedly pointed out by minority groups that, when such strategies were used as a basis for making employment decisions, they did not provide each person with an equal opportunity of being selected. It was pointed out that traditional strategies of selection were essentially fair to an individual within a group or class but unfair to the group or class as a whole.

The Quota Model

The arena of fairness in testing was moved to the courts and, in response to the first criticism noted above, a quota model was brought into play. The quota model was invoked on numerous occasions in the courts when a clear history of past discrimination could be shown. Essentially, as a selection strategy, it requires that a given number of all new applicants selected be from the group or class that was heretofore discriminated against.

The quota model placed great emphasis on a qualitative evaluation of past history as it related to equal opportunity. The model defined fairness in terms of a qualitative judgement about the history of employment practices. For example, if this model was applied to an employment selection situation in which it was shown that discrimination against women had been a command practice for many years, a fair solution may mandate that

for every male hired a female must also be hired. This solution would apply only to females who were at least minimally qualified and would be discontinued once an acceptable ratio of the work force was female. This model would be highly regarded by members of a least-favored group who were concerned about the group's welfare.

The Equal Risk Model

As to the second criticism, that traditional selection strategies did not afford each person an equal opportunity of being selected, R. M. Guion (1966) suggested the following definition of fairness.

> Unfair discrimination exists when persons with equal probabilities of success on the job have unequal probabilities of being hired for the job.[6]

In 1971, Einhorn and Bass quantified the above probabilities in a model referred to as the equal risks model of fairness. This model keyed in on an applicant's test score X and his probability of being successful on the job. The shift of emphasis to a person's test score X was a marked departure from the regression model's emphasis on a person's predicted score Y'. It also emphasized outcomes that were not considered by the regression model.

For any given employment decision regarding selection, there are four possible outcomes.

1. An applicant is selected who will be successful on the job.
2. An applicant is rejected who would have been successful on the job had he been selected.
3. An applicant is rejected who would have been unsuccessful on the job had he been selected.
4. An applicant is selected who will not be successful on the job.

These outcomes are shown in figure 2. The first and third outcomes are desirable and the second and fourth outcomes are undesirable. The equal risk model attempts to equalize the chances of an applicant's being accorded outcome two and an employer's being accorded outcome four.

Possible Outcomes

	Rejected	Selected
Would Succeed	Outcome 2	Outcome 1
Would Not Succeed	Outcome 3	Outcome 4

Figure 2

The Equal Risk model requires that the predictor cut score Xc be determined for each group or class in a way that will accord an equal probability of being selected for applicants within each group or class. It attempts to maximize both the employer's desire to obtain personnel who are most likely to be successful on the job and the candidate's desire to be employed. Such an attempt will afford some risk to the desires of both the employer and the candidate for employment. The equal risk model as a selection strategy attempts to equalize the risk for each party. In this sense, it may be regarded as a compromise position.

Code (1973) gave the following example of a hypothetical employment situation to illustrate how the cutoff points would be adjusted when the equal risks model is used.

> For example, suppose a selector is willing to hire all applicants with at least a 70% chance of success (or a 30% risk) as gauged by the predictor variables used. Then the predictor cutoff is chosen at the point at which the criterion pass point (Yp) is approximately one-half the standard error of estimate ($\sigma_{y.x}$) below the predicted criterion (y) since about 70% of the cases fall above minus one-half standard deviation in a normal distribution. [7]

Thus, if this model were applied to the employment situation above, the cutoff point would have to be adjusted to accept all job contenders who demonstrate a 70% chance of being successful and all those candidates who exceeded the cutoff would be equally considered for hire.

The Regression Model

In 1968, Anne T. Cleary reaffirmed the use of a regression model. Cleary performed separate regressions of Y on X (Y being a criterion of successful college achievement and X being scores on the Scholastic Aptitude Test (SAT)) for a group of white students and a group of black students. This was done at several integrated colleges. At each college, the regression lines for blacks and whites were compared for equality of slopes and equality of intercepts. Cleary points out that the regression model is a fair strategy for selection when the slopes and the intercepts of regression lines are equal for each group or class for which a prediction is to be made. Only one college was found in which the regression lines for blacks and whites were significantly different, and at this college, the black students scores were overpredicted. Thus, if a common or a white regression line was used as a basis for selecting students for admission to this college, there would be bias in favor of admitting blacks while whites would be penalized.

Cleary defined test bias as follows:

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression is consistently too high or too low for members of the subgroup.[8]

Under such conditions, a test would be considered as unfair. (Note: When the intercepts are not equal, consistent nonzero errors of prediction are being made.)

The Constant Ratio Model

The third criticism leveled against traditional selection strategies was that their use may result in fairness being accorded to individuals within a group but not to the group as a whole. This issue was not addressed until 1971. At this time, Thorndike pointed out that a strategy of selection considered fair by the regression model for individual members in both a minority and a majority group could in actuality be unfair for the minority group as a whole. He keyed in on the predictor scores X and demonstrated that, when the majority and minority groups differed appreciably in mean test scores on X, it would produce a situation that was fair to individual members of the group scoring lower on the test but unfair to the lower scoring group as a whole.

Thorndike's solution to this problem was a constant ratio model in which members from each group were selected in accordance with the proportion of a group's selected candidates for the job (Sc) to their original contenders for the job (Oc). For example, consider a job situation in which there were 40 contenders for a job - 20 males and 20 females. Of the 40 original contenders, 12 males and 8 females met or exceeded the minimum qualifications for the job and became candidates. For the constant ratio model, a state of fairness exists when minimum performance scores are set at levels that will qualify applicants in each group in proportion to So/Oc. If the probabilities for outcomes shown in figure 2 were quantified, the constant ratio model could be expressed as a ratio of (outcome one + outcome four)/ (Outcome one + outcome two). Peterson and Novick (1976) described the probability outcomes for six different models of fairness and compared them using a similar but more detailed notation. In our contrived example, the proportion of male contenders to male candidates was 12 to 20 or 60%, and the proportion of female contenders to female candidates was 8 to 20 or 40% (3/5 = 60% and 2/5 = 40%). Thus, if the constant ratio model were used, the criterion performance score would be adjusted to include a ratio of 60% males to 40% females, or for every 3 males hired, 2 females would be hired. For many current job situations, this model of fairness would most likely be preferred by women.

The Subjective Regression Model

In 1971, R. B. Darlington considered four previous models of fairness and recommended a common sense approach, which he called the subjective regression model. Darlington maintained that optimum treatment of cultural factors in test construction or test selection cannot be reduced to completely mechanical procedures. He argued that a subjective policy-level decision must be made when conflict arises between the two goals of maximizing a test's validity and minimizing a test's discrimination against a given cultural group. Darlington's model suggests that a test is fair when a proper balance between performance considerations and social and cultural considerations is obtained.

The subjective regression model requires both a performance measure (Y) and a cultural measure (C) for which a value estimate (K) is made. Darlington maintains that, once this datum is available, an investigator can specify a linear function of Y and C, or (Y-KC) which maximizes the balance between performance and cultural considerations. For example, in the employment situation described for the constant ratio model above, there were 40 original contenders for the job and half of them were minimally qualified and therefore became candidates. Twelve of

the candidates were male and eight female. Let's assume further that the scores for the candidates were as follows: the four highest scores were made by males, the fifth highest was made by a female, and the two lowest scores were made by a male and a female who were both of Spanish-speaking origin. It is noted that all candidates for the job in question are minimally qualified. If the subjective regression model were used as a selection strategy for the above job situation, values for C and K would be needed.

It would be necessary to assign some value (c) that favors one culture group when compared with another. Let's assume that the female of Spanish-speaking origin was assigned a least favored score, from the standpoint of cultural differences, of zero, (C=0) on a zero to ten scale, and that her performance score was 70 (Y=70). Also assume that the highest male performance score was 90, or (Y=90); and his score based on cultural differences was a most favored score of 10 or (C=10). Finally, let's assume that the value estimate for K was .2.

If the subjective regression model was used as a selection strategy for the job situation described above, the highest scoring male (Y=90) and lowest scoring female of Spanish-speaking origin (Y=70) would have equal scores for selection purposes. The male's score would be Y-KC=90-(.2) (10) = 70 and the female's score would be Y-KC=70-(.2)(0) = 70. Both would have selection scores of 70 and be equally considered for the job.

In a recent article by Darlington (1976), the task of determining a value to be used for C and K was acknowledged as being very difficult to perform. In this article, Darlington reaffirmed his support for a rational selection model and introduced two new methods whereby cultural factors could be included in the selection strategy. This resulted in a change in his methodology but not in his basic point of view.

The Conditional Probability Model

The next model of fairness to be considered is the conditional probability model. This model was advocated by N. S. Cole (1973). The model is similar to the constant ratio model in that it addresses the issue of group fairness as well as individual fairness. It favors the applicant's point of view at the expense of the employer and offers a solution to the undesirable outcome in which an applicant is rejected who would have been successful on the job had he or she been selected.

It differs, however, from the constant ratio model in the basic formulation of minority and majority group selection practices. Under the constant ratio model, selection for each group was made in accordance with the proportion of selected candidates (Sc) to original contenders (Oc) or Sc/Oc. Under the conditional probability model, the basis for selection was described by Cole as follows:

> The basic principle of the conditional probability selection model is that, for both minority and majority groups whose members can achieve a satisfactory criterion score ($Y > Yp$), there should be the same probability of selection regardless of group membership.[9]

The importance of an equal probability of being selected is amplified when one considers the fact that, at the time a selection is made, only the predicted criterion Y is known for each candidate; at this time the candidate's true score Y is yet to be determined.

For the conditional probability model, selection is considered as fair when the predictor cut scores Xc on each group are chosen so that a constant probability of selection will be accorded all applicants who have potential of being successful if given the chance.

If the applicants with potential for success from the majority group are accorded a 70% probability of being selected, then the applicants with potential for success from a minority group must be accorded the same conditional probability.

Again, if the possible outcome shown in figure 2 were quantified, the conditional probability model could be expressed as a ratio of (outcome one) / (outcome one + outcome two).

If this model of fairness was used in our contrived employment situation described for the constant ratio model above, we could expect the following: First, for each group the predictor scores would be adjusted until an equal conditional probability is accorded each group; then all candidates would be given an equal chance of being selected regardless of group membership. Let's assume further that only three persons were to be hired. Under such constraints, we would expect the three persons to be hired to be a male, a female, and a person of Spanish-speaking origin.

The Equal Probability Model

In 1973, Linn described a model that was later referred to as the Equal Probability Model. This model keys in on the applicants' scores on the predictor scores X for each group, and considers the probability of success of all applicants who are selected. Again, referring to figure 2, it is noted that two outcomes are possible for applicants who are selected. Outcome one and outcome four. If the probability of these outcomes was quantified, the equal probability model could be expressed as a ratio of (outcome one) / (outcome one + outcome four). This ratio could be determined for each group, and when the ratio was equal to a fixed constant for each group, selection by this model would be regarded as fair.

Consider a job situation in which 20 males and 20 females were selected as candidates for a job. After being hired and working on the job, 12 males proved to be successful and 8 females were successful. Thus, 8 males and 12 females were unsuccessful. If the equal probability model was used as a selection strategy for this job situation, one would expect further hires to favor women being selected.

Recent Models

During the last two years (1975 - 1976), a number of new models have been discussed in the literature. The thrust of these new models will be briefly considered beginning with the Threshold Utility Model. This model was first suggested by Gross and Su (1975) and later by Peterson. Peterson and Novick (1976) discussed this model within the framework of a rational decision-making paradigm. This model was to serve as a means for maximizing expected utilities given different points of view as to what constitutes utility. What is of great value from the point of view of an employer may not be considered as being as valuable from the applicant's viewpoint and vice versa. The threshold model recommends that a summarization of values from various points of view be made and that the decision with the highest utility pay off be accepted as being essentially fair.

Peterson and Novick (1976) introduced four new models that were converse to models that had previously been discussed. These are (1) the converse equal risk model, (2) the converse constant ratio model, (3) the converse conditional probability model, and (4) the converse equal probability model. The converse models were not presented by Peterson and Novick as desirable selection strategies; the models were merely introduced to make the previous models coherent within the decision-making paradigm they have committed themselves to. In effect, the converse models were used to show that their counterparts were inconsistent with the rational decision-making paradigm.

Another noteworthy model recommended this year was one by Lindley (1976) called the Normal Ogive Model. This model attempts to quantify those crucial decisions that Darlington left to a common-sense rational judgement. Also noteworthy is the fact that a computer assisted data analysis (CADA) system is being developed to complement this selection strategy. The normal Ogive Model along with twelve other models and the sources that recommended them, are listed in figure 3.

The Current Solutions Being Advocated

Presently, there are three distinct schools of thought about how the puzzle of fair selection may be solved. These have been identified and interpreted as follows:

a. One school of thought holds that a quantitative mathematical approach be used and that solutions that are formulated remain within the constraints of a rational decision-making paradigm. Value is based on, and equated with, the utility of outcomes that result from decisions made; and the worth of a selection strategy is judged on the basis of the sum total of the utility it affords. The solutions advocated by Petersen and Novick can be identified with this line of thinking.

b. Another school of thought holds that a quantitative mathematical approach be used as a practical tool to arrive at solutions to problems that are spawned from human concerns. In essence, the final outcome justifies the process that is used. The solutions advocated by Sawyer, Cole, and Cole can be identified with this line of thinking.

c. Another school of thought holds that a strict mathematical/mechanical method will never be formulated to capture fairness within the overall selection process, and that, any mathematical formulation must be augmented by human judgement in areas where value is an issue. Solutions advocated by R. B. Darlington can be identified with this line of thinking.

Much could be said in favor of school (a). However, it is much too early to pass judgement on such an approach. In mankind's intellectual history, the use of paradigms has been both beneficial and detrimental. For example, Aristotle developed a form of logic that was confined to a syllogistic paradigm that was without doubt beneficial to mankind at one point in history but was later detrimental. The syllogistics paradigm was so well formulated that alternate approaches were not ever seriously considered until the 20th Century. Nevertheless, an approach that calls for clearness and precision is to be commended--be it syllogistic or rational decision paradigm.

As for school (b), it also has strengths and weaknesses. Two important tendencies that need to be considered about school (b) are its use of mathematics as a tool and its refusal to be stifled by formality. School (b) does not begin with a formal value commitment that is consistent across all concerns. However, if an imperfectly formulated process leads to a desirable outcome that satisfies human concerns about employment decisions, it must be accorded some worth. After all, many human concerns about the selection process have not been brought under a consistent, systematic value theory. The weakness of school (b) lies in its lack of formal structure while its strength is in its responsiveness to human concerns.

As for school (c), its strength would lie in its ability to respond to the here and now. If a consistent coherent mathematical means were available for use within the total selection process, by all means it should be used. However, as of now, this is not the case. For example, little, if any, attempt has been made to quantify utilities associated with the outcome in which a person's situation is not improved (figure 3). The psychological, social, economic, legal, and political utilities of this outcome stand as an awesome task for any school to evaluate.

In final analysis, the issue of fairness has not been resolved; however, it is being vigorously pursued with precision, concern, and common-sense.

Axiological Models of Fairness

Model	Recommended by
Regression Model	This model has traditionally been used as a selection procedure. It was reaffirmed by A. T. Cleary 1968, A. Anastasi 1968, Q. McNemar 1975.
Quota Model	Numerous court decisions within the last ten years.
Equal Risk Model	R. M. Guion 1966, H. J. Einhorn and A. R. Bass 1971.
Constant Ratio Model	R. L. Thorndike 1971
Subjective Regression Model	R. B. Darlington 1971
Conditional Probability Model	N. S. Cole 1973
The Equal Probability Model	Linn, 1973 (discussed by)
The Threshold Utility Model	Gross and Su, 1975 later Petersen
The Converse Equal Risk Model	Petersen and Novick 1976.
The Converse Constant Ratio Model	Petersen and Novick 1976.
The Converse Conditional Probability Model	Petersen and Novick 1976.
The Converse Equal Probability Model	Petersen and Novick 1976.
The Normal Ogive Utility Model	Lindley (1976) described by Novick and Petersen (1976).

Figure 3

FOOTNOTES

1 Public Law 88-352, as approved by the President on July 2, 1964; Civil Rights Act of 1964 with Explanation. Commerce Clearing HOuse, Inc., Chicago Illinois 1964, page 71.

2 Supreme Court of the United States. 401 US 424, Griggs et al versus Duke Power Company. No. 124 - October Term, 1970 (Decided March 8, 1971).

3 Subcommittee on Labor of the Committee on Labor and Public Welfare United States Senate; The Equal Employment Opportunity Act of 1972; U. S. Government Printing Office; Washington 1972. page 16.

4 Equal Employment Opportunity Coordinating Council; Uniform Guidelines on Employee Selection Procedures, Federal Register, Vol 41, No. 136 - Wednesday, July 14, 1976.

5 Equal Employment Opportunity Coordinating Council.

6 Guion, R., Employment tests and discriminatory living. Industrial Relations, 1966, 5, 20-37, page 26.

7 Cole, Nancy S., Bias in Selection, Journal of Educational Measurement, vol 10, No. 4., Winter, 1973, 237-255, page 239.

8 Cleary, Anne T., Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges; Journal of Educational Measurement, vol 5, No. 2, Summer, 1968; 105-124, page 115.

9 Cole, Nancy S., (1973) page 240.

OTHER REFERENCES

Anastasi, A., Psychological Testing, 3rd ed, New York; MacMillan 1968.

Cleary, T. A.; Humphreys, L. G.; Kendrick, S. A.; Wesman, A.; Educational Uses of Tests with Disadvantaged Students, American Psychologist, January 1975, 15-40.

Darlington, R. B., Another Look at "Cultural Fairness", Journal of Educational Measurement, volume 8, no. 2, Summer 1971, 71-82.

Darlington, R. B., A Defense of "Rational" Personnel Selection, and Two New Methods; Journal of Educational Measurement, volume 13, no. 1, Spring 1976; 43-52.

Einhorn, H. J. and Bass, A. R., Methodological Considerations Relevant to Discrimination in Employment Testing, Psychological Bulletin, 1971, 75; 261-269.

Gross, A. L., and Su, W.; Defining a "fair" or "unbiased" Selection Model: A Question of Utilities, Journal of Applied Psychology, 1975, 60, 345-351.

Guion, R. M., Open a New Window, Validities and Values in Psychological Measurement, American Psychologist, May 1974, 287-296.

Linn, R. L., Fair Test use in Selection, Review of Educational Research, 1973, 43, 139-161.

Linn, R. L., In Search of Fair Selection Procedures, Journal of Educational Measurement, volume 13, no. 1, Spring 1976, 53-58.

McNemar, Q., On So-Called Test Bias, American Psychologist, August 1975, 848-851.

Novick, M. R. and Petersen, N. S., Towards Equalizing Educational and Employment Opportunity, Journal of Educational Measurement, volume 13, no. 1, Spring 1976, 77-87.

Petersen, N. S. and Novick, M. R., An Evaluation of Some Models for Culture-Fair Selection, Journal of Educational Measurement, volume 13, no. 1, Spring 1976, 3-29.

Sawyer, R. L., Cole, N. S., and Cole, J. W. L., Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection, Journal of Educational Measurement, volume 13, no. 1, Spring 1976, 59-76.

Thorndike, R. O., Concepts of Culture-Fairness, Journal of Educational Measurement, volume 8, no. 2 , Summer 1971, 63-70.

DETERMINATION OF MINIMAL ACCEPTABLE PERFORMANCE LEVELS FOR CRITERION-REFERENCED, MULTIPLE-CHOICE TESTS

John B. Meredith, Jr.
Myron A. Robinson, Ph.D
Data-Design Laboratories
Norfolk, Virginia 23502

INTRODUCTION

One of the major problems confronting the designers of criterion-referenced, multiple-choice tests is the determination of passing scores. In general, passing scores for these tests are arbitrarily set at 80% to 90% success (Sanders, 1976; Shaycoft, 1976). This method for arbitrarily setting passing scores has the following disadvantages:

1. Nonconsideration of test item difficulties, and

2. Greater classification error, where true masters are classified as nonmasters and true nonmasters are classified as masters (Reichman and Oosterhoff, 1976).

This paper describes the application of an innovative method developed by Nedelsky (1954) for setting passing scores that minimizes these disadvantages. This method involves the definition of "minimally acceptable" performance and the derivation of a Minimal Acceptable Performance Level (MAPL) for the test. In addition, this paper presents a method for the evaluation of training, based upon criterion performance defined by the MAPL.

The MAPL is determined by using Subject Matter Experts (SMEs) to rate the difficulty of the alternatives to each item of the test for the minimally acceptable (just passing) examinee (Meskauskas, 1976). Item difficulty is criterion-referenced, since it is based upon item content rather than examinee performance (Smilansky, 1976). The SME ratings are used to establish criterion standards for each item, and the sum of these criterion standards are then used to establish the MAPL for each area of the test. In fact, this approach to the problem of determining passing scores provides one of the best estimates of the probability of classifying examinees into the correct mastery and nonmastery states (Reichman and Oosterhoff, 1976).

The MAPL technique was used in this study to determine the extent to which a training program produced equivalent proportions of minimally acceptable technicians in various areas of specialization.

METHOD

Nedelsky (1954) developed the MAPL technique as an "absolute standard" for evaluation of physics students on a departmental, multiple-choice comprehensive examination at the University of Chicago. The MAPL technique was validated by Taylor and Reid (1972), Bobula (1974), and Smilansky (1976). The use of this technique is dependent upon the assumption that SMEs can define alternative similarity as follows:

1. An alternative which a minimally acceptable examinee should recognize as incorrect is given a value of zero (0),
2. An alternative which a minimally acceptable examinee should not recognize as incorrect is given a value of two (2),
3. An alternative which is correct is given a value of two (2), and
4. All other alternatives are given values of one (1).

An example of the application of this MAPL technique, adapted from Bobula (1974), is exemplified by the faculty member who is teaching statistics and defines ability to recognize measures of central tendency as a component of basic statistical competency. He might write an item and assign the following values:

Value		Item
		"The most appropriate measure of central tendency for the reading performance scores of a large class is the"
	(2)	1. Mode
	(0)	2. Variance
Correct	(2)	3. Mean
	(2)	4. Median
	(0)	5. Standard deviation

When the SMEs rated this item for the minimally acceptable examinee, they decided that three alternatives (Mean, Median, and Mode) were eqully viable, and all were given values of two (2).

The Alternative Similarity Index (ASI) of an item is defined as follows:

$$ASI = \frac{\text{Value of correct alternative}}{\text{Sum of values of all alternatives}} = \frac{2}{6} = .33$$

The MAPL for each area of the test is the sum of the ASIs (rounded to the nearest integer). A passing score is given when the examinee's score is equal to or greater than the MAPL. The number of examinees achieving the MAPL is determined for each area of the test. From the numbers of examinees achieving the area MAPLs, the confidence interval about the mean number of examinees achieving this passing score is determined. Any area of the test where the number of examinees is below this interval indicates the possibility of undertraining. Conversely, any area where the number of examinees is above this interval indicates the possibility of overtraining.

FINDINGS

In this study, five SMEs evaluated a 360-item test to determine the MAPL for each of 12 areas of specialization. The area MAPLs were derived and are shown in Table 1.

Table 1. Area MAPLs

Area	No. of Items	MAPL
A	10	6
B	29	14
C	30	15
D	46	25
E	32	16
F	46	24
G	24	11
H	33	15
I	24	11
J	29	15
K	32	15
L	25	14

Using each area MAPL, the number of examinees achieving the MAPL, the number of examinees not achieving the MAPL, and the percentage of examinees achieving the MAPL were derived (Table 2).

Table 2. Area Criterion Performance

Area	No. of Examinees Achieving MAPL	No. of Examinees Not Achieving MAPL	Percentage of Examinees Achieving MAPL
A	217	57	79.2
B	259	15	94.5
C	217	57	79.2
D	202	72	73.7
E	201	73	73.4
F	176	98	64.2
G	192	82	70.1
H	185	89	67.5
I	219	55	79.9
J	113	161	41.2
K	170	104	62.0
L	161	113	58.8

The mean percentage of examinees achieving criterion performance across all areas was 70.3 with a standard deviation of 13.3. The .99 confidence interval about this mean is $58.4 \leq \mu \leq 82.2$. Two areas were found to lie outside of this confidence interval: Area B (94.5) and Area J (41.2). These results suggest the possibility of overtraining in Area B and undertraining in Area J. These indications must be reviewed by SMEs before recommending changes to the training program.

CONCLUSIONS

This paper has presented a methodology for making preliminary training assessments from criterion-referenced, multiple-choice tests. An innovative technique for obtaining criterion level information was described and applied to the evaluation of test results for purposes of training assessment. Further research into validating this technique for this type of application is recommended. At this time, however, it appears that the MAPL technique offers an efficient method for developing preliminary indications of overtraining and undertraining.

REFERENCES

1. Bobula, J. Minimum Pass Level Study. Report to the Faculty, Center for Educational Development, College of Medicine, University of Illinois, 25-31, 1974.

2. Meskauskas, J. A. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard Setting, Review of Educational Research, 46:133-158, 1976.

3. Nedelsky, L. Absolute Grading Standards for Objective Tests. Educational and Psychological Measurement, 14:3-19, 1954.

4. Reichman, S. L. & Oosterhoff, A. C. Strategy Guidelines for the Construction of Mastery Tests, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April, 1976.

5. Sanders, J. R. Alternatives for Achievement Testing, Educational Technology, 16:March, 1976.

6. Shaycoft, M. F. A Paradox in Setting Cutting Scores on Criterion-References Tests, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April, 1976.

7. Smilansky, J. and Guerin, R. O. Minimal Acceptable Performance Levels for Criterion-Referenced Multiple-Choice Examinations and Their Validation, Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, California, April, 1976.

8. Taylor, D. D. & Reid, J. C. Criterion-Referenced Evaluation of Student Performance, Journal of Medical Education, 47:970-971, 1972.

Discussion of a Study of Occupationally Transferable Skills[1]

Allen A. Wiant[2]

Under sponsorship of the Education and Work Group of the National Institute of Education, The Center for Vocational Education at Ohio State University is working to identify more clearly the skills and abilities that schools can teach to prepare individuals, not only for a job, but for the series of jobs that will be their careers. This concern grows out of the work experience of the average American today, in which job mobility--both voluntary and involuntary--is apparently an increasing phenomenon. Under these conditions, future education and work programs will have to concern themselves with improving the ability of individuals to make career changes.

We are grateful for your attendance here this afternoon because we need your reactions to what we are doing and what we think we are finding. Because of your special expertise and concern for performance and measurement, we are particularly anxious to receive your input. Accordingly, this is intended to be primarily a discussion of the issues rather than a presentation.

Our current activity is an initial step in the direction of determining what schools can do to better prepare individuals for work careers characterized by change. We are attempting to assess the state of knowledge and to define appropriate next steps.

Thus, one objective of the present study is to identify the range of skills and personal characteristics which might be considered occupationally transferable. These might include basic cognitive skills, technical literacy skills, problem assessment, decision-making and judgmental skills, skills related to the manipulation of tools and machinery, multiple-use technical skills, and other individual attributes and characteristics.

A second objective is to identify and describe existing sources of occupational mobility data, highlighting those features that may support further R&D needs on job transferable skills and occupational adaptability.

A third objective of the current effort is to identify job classification schemes potentially useful for describing the skills, experiences, and other attributes contributing to occupational adaptability, and for research and development to improve occupational adaptability.

---

[1]Introductory remarks at a session of the annual conference of the Military Testing Association, Gulf Shores, Alabama, October 20, 1976.

[2]Mr. Wiant is a Research Specialist at The Center for Vocational Education, The Ohio State University, Columbus, Ohio.

A fourth objective of this effort is to hold exploratory discussions of the transferability of skills with those who are responsible for the selection and promotion of employees. This has involved a series of small workshops with businessmen and others to: (a) identify the types of transferable skills which are required in their work settings, (b) assess the importance of transferable skills in qualifying for the jobs and promotions over which they have control, and (c) assess the role of educational institutions in providing training for the development of transferable skills and the updating of these skills and abilities in adults who face occupational changes throughout their working life.

Some of the results of these efforts appear to be coming together. For example, James W. Altman, in an unpublished paper written to describe types of transferable skills, considers skills to be hypothetical constructs derived from observation of task performances, rather than operational realities. In his paper, no distinction is made between skills which are transferable and others which are not; all are presumed transferable, but their effective transfer depends on factors such as environmental context and individual needs. From our discussions with personnel and training officers in business and industry, we are compiling lists of skills and individual characteristics which they feel are transferable and essential for successful occupational adaptability. While no attempt has been made to collect this data in a scientific manner, the consensus appears to be compatible with the Altman view.

Following is a compilation, to date, of the skills, attributes, and attitudes which we have been told are the "things" which are occupationally transferable (the category headings and item assignments are arbitrary).

Aptitudinal/Intellectual

reading and interpreting
communication (writing)
public speaking
investigative thinking
abstract thinking
creativity
artistry
computing
analyzing
problem recognition
trouble shooting
problem solving
developing options
evaluating
synthesizing
integrating diverse ideas
organizing
logical decision making
planning and layout
forecasting
accounting/bookkeeping
budgeting
general clerical skills (esp. typing)
tool skills (mechanical)
multiple activity
familiarity with other jobs
using available resources
working efficiently
understanding people (others)
understanding self
ability to apply knowledge/experiences in different contexts

Interpersonal	Attitudinal
communicating (oral)	desire/motivation
supervising	adaptivity
cultivating cooperation	flexibility
delegating authority	enthusiasm
empathizing	acceptance of responsibility
counselling	stability
encouraging/motivating	objectivity
relating to others	acceptance of work
getting along with others	acceptance of direction
working with others	acceptance of risk
compromising	good work habits
leading	patience
persuading/selling	perseverance
listening	desire to learn
	self-discipline
	appreciating others
	tolerance
	self-confidence
	integrity
	loyalty

There has been remarkable consensus concerning the nature of these items, but perhaps this is simply a reflection of conventional wisdom. In any event, these findings, taken together with the perceptions of others with a more theoretical orientation, lead us at present to a number of tentative conclusions, including the following:

1. "Transferable skills" seem to have two characteristics:

   First, they are abstract elements of human ability/activity which are common to many life and work situations (e.g., decision making, planning, making use of available resources).

   Second, in the context of a particular work or life situation, they are involved in the performance of concrete activities.

   Each context in which a skill element is employed supplies a unique dimension to the demonstration of that skill.

   Thus, a transferable skill is like a general purpose tool, useful in a variety of applications. The user must have facility in the use of the tool, as well as knowledge/understanding of the application, in order for the tool to be truly useful.

2. Employers frequently affirm that individuals need only possess such basic skill elements in order to obtain employment, and that they will train their employees in the application of these skills in their own employment context. Others deny this claim, pointing out that those who get hired are those with job specific skills (i.e., performance and knowledge developed for a specific context).

3. There is disagreement among employers as to their ability to make a reasonable assessment of a job applicant's basic skills using presently available means. At best, however, the means are subjective and proxy indications are frequently the only data available (e.g., school attendance record as a predictor of job dependability).

4. For the above reasons (2 and 3), it appears that basic skills are presently regarded as being <u>necessary</u> but not <u>sufficient</u> for obtaining employment. Following employment, an employee's possession of these skills is typically assessed by observing specific task performances which are believed to require their use. These skills are therefore necessary for the maintenance of employment and employability, particularly the attitudinal and interpersonal skills.

Another finding of the project to date is that very little is known, based upon research, concerning the personality characteristics of occupationally mobile persons. Neither have we found data bases containing information on job change which appear to offer much potential for examining the basic phenomena of skill transfer or for identifying occupationally transferable skills.

There has been much attention given in this conference to the use of tasks as the basic unit for training and performance assessment. We share the conviction that the job task is the most appropriate unit of analysis for job training, and continue to work extensively in this area. However, we also hold the view that whereas we know how to do a good job of task enumeration, we know less about a technology for analyzing job tasks to describe the skill requirements of these tasks. Until we have such a technology, task analysis cannot be depended upon to determine skill requirements. In the interim, we will need other means.

We are highly gratified by the degree of interest indicated by the attendance at this session, and hope that you will have thoughts and reactions to share with us now concerning the notion of occupationally transferable skills. In particular, if you know of programs which recognize and attempt to make use of this notion, or of either data or studies which might shed light on this question, we would like to hear of them.

THE JOINT SERVICES APPROACH TO ENLISTMENT

PROCESSING

BY

WILLIAM P. ACKER, BG, USAF

DEPUTY COMMANDING GENERAL

U.S. MILITARY ENLISTMENT PROCESSING COMMAND

Admiral Wilson , Ladies and Gentlemen. I am quite pleased to have this opportunity to review MEPCOM operations and common problems with such a distinguished assemblage of professionals. It is always a pleasure for me to work with joint services' organizations -- working on common problems that cut across service lines. You know, offtimes, we in the military have a tendency to stick to our own little niches, forgetting that our concerns are frequently shared by many other people in various agencies. We do work together on the important things; such as combat. Most of us also enter the military for the first time under a common service program -- the Armed Forces Entrance and Examining Stations (AFEES).

Since July of this year these AFEES have been realigned into a separate joint service command; The Military Enlistment Processing Command (MEPCOM), headquartered at Fort Sheridan, IL.

The Commanding General of MEPCOM (General Forrester) is also the Commanding General of the U.S. Army Recruiting Service, USAREC. Below this level of command, the MEPCOM organization is joint service staffed - with 130 persons assigned at the Headquarters, 30 each at one of our three sectors, and 2760 joint services personnel assigned to the 66 AFEES.

Organization of MEPCOM brought together the old Production Management Directorate of USAREC and the Armed Forces Vocational Testing Group (AFVTG), Randolph AFB, which was the single manager for the DOD High School Testing Program.

Under the pre-existing system, too often the same individual at varying levels of commands which was responsible for meeting enlistment quotas for the Army also was charged with the quality of input. Under the MEPCOM arrangement, we have disassociated the responsibilities for managing the quantity of accessions from the management of insuring the quality of input. That is only one example of the advantages associated with creation of a separate command for enlistment processing. There are many others. For example: there was a savings of 2.3 million dollars in forms reduction alone -- even before MEPCOM was chartered as a separate Command. We were also able to free service dedicated ASVAB test administrators back to the recruiting commands.

Although it is much too early to assess overall effectiveness of MEPCOM, several of our programs are already demonstrating a payoff. We have instituted a "RED CARPET" indoctrination for all new recruits. In terms of receptiveness, attitude change and tension reduction, this program has proven to be a very popular success.

With introduction of the common service ASVAB in January of this year, MEPCOM now manages the single largest standardized testing program in the world. Between 1 July of 1976 and the end of next June, we

expect to test some 1.3 million recruit applicants with the production versions of the test, and around 1.6 million high school students with the institutional edition of ASVAB -- in over 16 thousand high schools. It is quite apparent, that as more students are tested in the high schools, fewer enlistees will be required to be tested at the AFEES and MET sites. Already, we test approximately one out of every four 12th graders in the country.

As with any program of this size, all is not sweetness and light. Some Congressmen and organizations such as the American Civil Liberties Union are questioning DOD sponsorship of a testing program directed to the high schools. The common claim is that we are transforming high school counselors into pseudo-recruiters. A peculiar outgrowth of these concerns (and as behavioral scientists, I know you will see the implications in what I am going to read) is the requirement that an "ERRATA," as agreed to between Congressman Mosher (R-Ohio) and OSD (M&RA), be affixed to each document communicating ASVAB information to high school students, parents and counselors. Let me read this errata to you:

IMPORTANT INFORMATION

PURPOSE

The Armed Services Vocational Aptitude Battery (ASVAB) is used by the Armed Forces for recruiting purposes and by school counselors for vocational guidance counseling. The ASVAB's ability for determining civilian job skills has not yet been proven.

DISTRIBUTION OF TEST SCORES

Test scores are provided to the school counselor, who in turn furnishes them to the student, and are provided to the recruiting services of the Armed Forces and the Coast Guard.

RETENTION OF INFORMATION

Information about individuals who have taken the ASVAB will be maintained on a computer tape and on microfiche records for not more than two years. Test scores and demographic data required for research purposes will be kept for a longer period of time; however, personal identity information: name, Social Security number, street address and telephone number will be erased from research records after two years.

RELEASE OF TEST INFORMATION

Information on an individual's identity and the related test information will not be released to any agency outside of the Department of Defense except to the U.S. Coast Guard, which is under the jurisdiction of the Department of Transportation, and the student's school system. Information on the individual and the related test scores will not be used for any purpose other than for recruiting by the Armed Forces, vocational counseling in the school system and for research on test development and personnel measurement. Information on test scores provided to school systems are to be handled and disposed of in accordance with the policies of the governing state or local school system.

Of course, most of the provisions contained in this mandate -- which has the clout of a legal contract -- represent actions that were already being taken. I believe there are two provisions in the agreement that are of special interest to this group:

-First is the statement that; "The ASVAB's ability for determining civilian job skills has not yet been proven." This statement really means that ASVAB has not yet been validated against job performance at the journeyman level in civilian occupations. This is true, both in the military and in the civilian sectors. I am sure you are all aware of the problems involved in obtaining good, uncontaminated criterion measures of job performance. The statement also points out the need for clean data relating the content of technical training to job task composition and for equating service jobs to civilian equivalencies. I find it interesting that the technical subtleties of these statements are far too sophisticated for comprehension at the level of the high school student.

-Secondly, I think you all recognize the import to R&D efforts of slapping a two year retainability restriction on basic test data. We can't very well run an extensive longitudinal study under such restrictions.

I think we have reached the point of enough said about the Mosher Agreement. Please see me later if you are interested in the many actions we have had to take in order to insure full compliance with the intent of this new requirement.

In theory, the Defense Manpower Policy Council's decision to require a common service qualification and classification test battery across all services had the advantage of concentrating research efforts of the behavioral research laboratories around a single instrument. I can think of several areas where we could definitely benefit from your ideas, research conclusions and past experiences. Let me enumerate some of these problems. Please let us know if you have any ideas that we could use. After all, the experience assembled in this room represents the best available in the world in the use of large standardized testing program.

We have had, and continue to have, problems across the educational community with total testing time. Although total announced time for the institutional version of ASVAB is 2 hours and 45 minutes (which includes 30 minutes of administrative time), we are finding that many test situations -- particularly in large school settings -- extend beyond 3 hours. Optimum time is set between arrival of that last bus in the morning and beginning of the first lunch period. We need a concentrated effort to define exactly what type measures produce maximum predictive validity for use in the high schools, and which ones of these are absolutely essential for non-prior service recruit classification. It may be that the requirements of the two programs are different. In any event, total test time in the schools should be around 2 hours. Our kids are subjected to too much testing and academic "down time" as it is.

Another, quite serious, problem that we continuously run into with our production tests is that of test compromise. At the end of September we had 28 cases of test loss, and 36 instances of test compromise. Obviously, the underlying motivation behind the vast majority of these occurrences is a desire to obtain a higher passing score on the test. This is the opposite of our problems just a few years ago when many draftees purposely tried to achieve a non-qualifying score. Things have changed to the point where tests have been compromised under the threat of a .38. We actually had a test administrator held up on the east coast under armed robbery. Needless to say, he surrendered the test and scoring key without argument.

Although we have drastically tightened up on security procedures, we are still searching for any supplemental hints, techniques or psychometric guides that would help reduce the probability of occurrence. I do not believe that reissuing a completely new series of tests, in multiple versions, every 2 years is the answer. Perhaps the ultimate solution lies in the future, in the form of Computer Aided Testing. At least with computers we wouldn't have to worry about test storage, shipment, answer keys and booklets.

To move into another problem area, I notice that our AFEES Test Administrators presently have to translate ASVAB raw scores into 24 service specific conversions -- not counting the conversions required for the high school program. Moreover, each service seems to have its own conversion metric; with the Navy using a standard scale which is different from that of the Army, and both of them different from the percentile scale used by the Air Force. With the advent of the common service ASVAB, I would like to think that we could all agree, at least, upon a common metric for use across all services; and, perhaps, even a common normative base -- may be even based on the representative high school norms now available. I ask you to give the idea some thought.

You know, for some years now, the popular criterion used for validation of our various service tests has been final course grades (FCG) in our structured technical training courses. The FCG is depended upon almost exclusively as the measure of how successful (or unsuccessful) are our test measures. I have no quarrel with this procedure, but feel that there is some merit in looking at other criteria which are both available and important. Specifically, I would like to suggest that we look at the factor of attrition in each study where initial acquisition test scores are correlated against performance in the technical training environment.

Both the Air Force (Dempsey & Fast, 1976) and the Navy (Neumann & Abrahams, 1976) have had some success in using non-cognitive measures to predict recruit turnover in the Academy setting. Given a rather stable trainee dropout rate around 13% in Class "A" type courses, any increase in prediction efficiency would result in tremendous savings to the services. Even a 1% decrease in turnover would literally save hundreds of thousands of dollars per year. Maximum utilization of our constantly decreasing personnel resources, increased career retention and release of support personnel to the combat unit are all perceived as viable outgrowths of successful research along these lines.

I do know that the Air Force Military Personnel Center at Randolph, with the concurrence of the joint services ASVAB Working Group, is pilot testing a variety of such instruments. I am looking forward to results of the study.

To us in the AFEES business, it frequently seems that the recruiting services ship all prospective recruits for examining, regardless of their potential to read and write, let alone "pass" the ASVAB. We usually get our first clue when it is discovered the youngster cannot fill out the medical history (SF 88), or has difficulty completing the header information on the ASVAB answer sheets. What is really needed is a well standardized DOD wide Enlistment Screening Test (EST) which reduces to an absolute minimum the number of applicants sent to the AFEES whose probability of qualifying for enlistment is virtually nil. Once again, the joint services ASVAB Working Group is assessing various approaches to the problem,

and will hopefully agree upon a common EST. I see little to be gained by expending monies to transport and process applicants who could quite easily be screened out at the level of the recruiting stations.

Beginning in January of this year, we began to obtain ASVAB test scores on all enlisted personnel. Around January of 1978, we should begin to make plans to plug the ASVAB into out-processing counseling for all personnel being discharged from the services. At the present time, most base education and counseling offices use the General Aptitude test Battery (GATB), sponsored by the Department of Labor, as one of the primary counseling instruments for counseling enlisted personnel concerning civilian job opportunities. Working in close cooperation with the Department of Labor and the U.S. Office of Education, the old Armed Forces Vocational Testing Group assembled common service job descriptions -- related to their civilian counterparts in the Dictionary of Occupational Titles (DOT) -- and published these common job sheets in the DOD Military/Civilian Occupational Source Book. With the availability of these data coupled with the results of ongoing research which related ASVAB test performance to specific job profiles in the Source Book, we will have the capability to supplement GATB information with the ASVAB. I do not regard the ASVAB as a replacement -- or even a substitute for GATB. Both test batteries were born out of differing requirements, measure different attributes and satisfy diverse needs. We are, and will continue to, work closely with staff members in Labor to insure that the test batteries compliment each other, and that an open medium for the exchange of information remains open.

I do believe that, when a comprehensive aptitude battery reaches the point where it is useful and valid for both inprocessing and outprocessing placement and counseling, it then meets its level of optimum cost-effectiveness.

At the point in time that the institutional ASVAB (1972-1973) was introduced to the high schools under single DOD management, the test fortunately coincided with nation-wide emphasis on the need for career education. Without this emphasis, it is doubtful if over 1.2 million high school students in some 16 thousand schools would have opted for the test. Career education, as a program, is continuing as a strong force, and is expected to gain momentum as counselors across the country become more concerned with acquainting their students with occupational reality. I think the services would be well advised to adopt some of the concepts of career education to their own in-service operations. We place heavy emphasis upon the benefits of service related career education in our recruiting literature, but frequently ignore such admonitions when dealing with our own people. For example, in spite of repeated assurrance that a recuit is guaranteed school assignment prior to enlistment, almost one out of every five non-prior service enlistees does not successfully complete that first technical training course. What happens to these people? What do they tell their buddies back in their hometowns? Do we provide these in-service "drop-outs" adequate career counseling? I can see many areas where we need to improve our humanity. Under the umbrella of the all-volunteer force, we need to review all of our personnel programs

to insure that our vital personnel resources are treated as individuals -- with dignity -- and not as collective members of the herd.

We work too hard, and it is extremely costly, to acquire each and every single recruit, to end up treating these people as if they were stamped from the same mold. I know these are platitudinous statements, but I feel strongly about the subject. Attitudes must change, and programs must change, if we are to sustain the vitality of our individual responsiveness under the all-volunteer force.

In 1973 there were many, many people placing bets that the services would not be able to attract a sufficient number of quality recruits to sustain an all-volunteer service. Time -- and hard work -- have proved these dooms' day prophesiers to be wrong. Although dips in the nation's economy have definetely contributed to this success, I am convinced that the high school ASVAB was an equal contributor. We do know that, in those areas where the institutional program is used and pushed strongly by all the services, recruiting services in those same areas consistently seem to meet their quotas; recruiters swear by use of the ASVAB prospect lists; and relationships with the educational community are good. Perhaps such success is a function of close liaison between the local recruiter elements. It does point out the continuing need to work together, for the good of all the services.

I appreciate your attention and interest. At this time I would like to open the floor to any questions you may have regarding MEPCOM; our organization, mission, procedures, or plans. Thank you.

DEUTSCHER BUNDESWEHR-VERBAND E.V.
THE FEDERAL ARMED FORCES ASSOCIATION
-Federal Executive Board-

5300 Bonn-Bad Godesberg, 18 oct 1976
Südstraße 123, Tel (02221) 31 oo 21

MILITARY TESTING ASSOCIATION
October 18 - 22, 1976
Gulf State Park Resort,
Gulf Shores, Alabama

Presentation
Thursday
October 21
0900

PROBLEMS OF BILLETS CLASSIFICATION AS SEEN BY THE FEDERAL ARMED FORCES ASSOCIATION (DBwV)

Address

Colonel Hans-Erich Seuberlich
Chairman Army Section
Federal Executive Board (DBwV)

Bonn
West Germany

Problems of Position Assessment -
As seen by the Deutscher Bundeswehr-Verband

1 - Introduction

Permit me, gentlemen, to first introduce to you briefly the Deutscher Bundeswehr-Verband - the German Federal Armed Forces Association - which I have the privilege of representing here.

This introduction before the MTA seems of special importance to me because the MTA sees itself as an institution engaged in experiments. And the Deutscher Bundeswehr-Verband - one might say - at the time of its founding really was a great and novel experiment in German military history. I will return briefly to this subject later. In order to be able to present details on it I would like to suggest that this be done either in discussion or that an additional working group be formed in which those interested may deal with the subject. But now to our brief introduction of the "Association", as I'll call it from now on.

When the Federal Armed Forces - the Bundeswehr - were formed in the fifties, Federal Parliament - the Bundestag - and legislators stipulated that the armed forces should be firmly integrated in state and society. For this purpose the image of the "Citizen in Uniform" was created. In this context, the constitutional freedom of association - open to one and all - for the first time also applied unrestrictedly to servicemen.

This basic right provided the base for founding our Association. In 1956, that is exactly twenty years ago, fifty-five soldiers founded this novel professional representation of servicemen in the garrison of Munster. Field-grade officers, company commanders, platoon leaders, senior and junior NCO's and privates were among the founding fathers. A battalion commander was elected first chairman. This great venture, then without parallel in German military history, has become a tremendous success. Today - twenty years later - the Association numbers 203,000 members of all grades and ranks. Our highest-ranking soldier, the Chief of Staff of the Armed Forces, is a member of the Association as well as numerous draftees - or conscripts.

Other than the armed forces, the Association has a democratic structure. Regardless of rank and position every member has but one voice, and all boards up to the Federal Board are elected democratically in secret balloting. Not weight of rank but sound arguments count.

What is the secret of success of this association, which is independent of the "employer"? I would like to name four decisive basic principles - adopted by the Association at its founding and adhered to strictly.

(1) Representation not only of economic and social, but also of the ideal interests of the servicemen.

(2) Non-interference in military duty matters, which also includes a decided acknowledgment of the indispensable principle of command and obedience.

(3) Rejection of labor conflict measures - and especially that of striking - as means to enforce association demands.

(4) Strict neutrality regarding religion and political parties.

With emergence of the Association, an epochal development was launched: For the first time, servicemen were given a constitution-based platform for own activities in the socio-political field. Freedom of association - or freedom of coalition as we call it - in the Federal Republic of Germany in particular guarantees the right of forming associations to safeguard and promote labor and economic conditions. With this I am now turning to the main subject of my presentation.

2 - The Bundeswehr - Part of the Public Service

The service laws of the public service in the Federal Republic of Germany encompass the general fields of status and career matters, pay and benefits. For civil servants, judges, teachers and soldiers these matters are governed by federal and state laws, while employees and workers - that is personnel receiving wages - are provided for by collective bargaining. Shaping of public service laws is influenced by the federal government, state governments, the respective parliaments and also by the organizations representing the various professional groups. These are mainly the large German Civil Servants' Federation, various unions and some smaller associations. Without their association, our servicemen would not be in the position to have their numerous service pecul-

iarities - which have to be taken into account in the fields of career, pay and benefit legislation - represented in the same manner as, for instance, teachers, judges or the police. This would result in a series of disadvantages for - or even discriminations against - the servicemen because their "employer", the Federal Minister of Defense - cannot by himself prevent - in our system of parliamentary democracy - the unions or other large organizations from pushing through decisive advantages for the professional groups they represent in legislation of the various levels. Add to this that the armed forces quite naturally effect an especially high rise of costs at times of income improvements because of their large number of personnel. Since expenditures for personnel amount to almost three quarters of the operating costs of the Bundeswehr anyhow, this important factor often serves as reason or excuse for excluding the soldier from participation in financial improvements. Thus, a number of discriminations in the field of pay legislation at least in part can be attributed to this fact. To name the most important ones:

- Servicemen do not have set or guaranteed duty hours, whereas everywhere else in public service the 40-hour week has been in effect for a long time already.
- Servicemen do not receive remuneration for additional work or duty - that is overtime pay - since it is argued that where there is no agreed upon limit to duty hours there can be no overtime.

  In this context I may point out that the reduction of the 42-hour week to a 40-hour week in 1974 equalled a 5 per cent pay raise.

- Soldiers do not receive - contrary to the Federal Border Guard - the so-called police-allowance amounting to DM 120.00 a month, even though the Federal Border Guard in many respects exercises comparable functions. But in contrast to the small organization of the Federal Border Guard - about 20,000 men - this allowance would cost the Bundeswehr more than ten times as much - namely 250 million - per annum.
- Fourth, and this brings me almost to the heart of our subject: While in the Federal Border Guard, for instance, the squad leader is given the classification of a Staff Sergeant (Feldwebel, OR-6), the underrating of the Bundeswehr's squad leader in some especially critical or sensitive fields - as the position of squad leader in an armored infantry squad - represents a striking example of discrimination, even though additional costs in this case quite contrary to official statements would not weigh that heavily. According to my own calculations, they would only amount to approximately 2.3 million DM per year.

Against the background of these circumstances, the two most important demands in the field of pay legislation have to be seen which were passed at the Ninth General Meeting of our Association in 1973, namely:

(1) The demand for a function-oriented pay for servicemen, that is pay according to position criteria, because the present system no longer takes specific functions adequately into account.

(2) The demand for an objective position assessment and rating

in all parts of the Bundeswehr as a prerequisite for function-oriented pay - because only with the aid of such assessment the various functions can be defined with adequate accuracy - as well as a prerequisite for meeting the general demand for realization of the ability and performance - or efficiency - principle: You can only pay according to ability and performance where a clearly defined yardstick of position criteria exists.

Today, this is more imperative than ever in the whole field of our public service, whose performance is being increasingly and thoughtlessly questioned. Our servicemen cannot exclude themselves from this demand. Of course, quite often it is even more difficult to measure their performance objectively than it already is the case with other parts of the public service. That is why conception and implementation of a method for objective position assessment has become especially important.

In 1975, the "Principle of Function-Oriented Pay" was written into the German Public Service Pay Law for the first time - and thus the demand of the Bundeswehr-Verband met at least theoretically. The passage in the law reads as follows:

"The functions of civil servants, judges and soldiers must be assessed objectively according to the criteria connected therewith and (to be) assigned to offices (positions). These offices are to be assigned to pay groups according to their value and under consideration of the common interests of all principals (employers)."

3 - Position Assessment in the Public Service of the Federal Republic of Germany

Behind the problem of position assessment lies the wish for equity in pay. This is an understandable and old wish. The difficulties in regard to its realization and the multitude of attempts to achieve it are reflected in the history of German pay legislation. Shortly after the turn of the century the first German pay laws by modern standards came into being:

- The Imperial Pay Law of 1909 had 180 pay classes and thus reflected the number of functions existing at that time.
- The Pay Law of 1920 condensed the existing 180 pay classes to a mere 20 pay groups in a generalizing manner. At that time, too, the first uniform principles on career matters were formulated and linked with the system of the so-called "position formula" (Stellenschluessel) - an almost universal rule of position distribution by ratio and applied to all parts of the public service. This led to a strongly generalizing pay system with similar weaknesses and inadequacies as we experience today - and which I had the honor to speak on briefly before the MTA in 1973.
- Beginning in 1927 legislators tried to create a more function-oriented and diversified pay system by increasing the existing one to 47 pay groups. Now, assessment (and classification of the individual position was to be basis for promotion.
- In the Pay Law of 1969, finally, the number of pay groups was drastically reduced again - from 47 to 27, thereby almost cutting them in half.

This meant a departure from strong differentiation with the consideration of peculiarities inherent to certain individual functions. Simplification was the aim of this measure. The attempt to achieve it in conjunction with a proportionally equal distribution of positions permitting promotion on all federal as well as state levels proved a failure. The lump-sum setting of upper limits by the legislators could not even begin to consider adequately the numerous differences in functions of higher value in the various occupational fields. Manifold new injustice had to be the inevitable consequence.

That is why the Bundestag - Parliament - already two years later requested the Federal Government to fill the principles of position assessment with greater detail and to speed work on a classification system which paid greater attention to position criteria. With this request parliament tried to finally stop diverging developments in various states and parts of the public service. Thus, for instance, the city-state of Hamburg had called for a reassessment of one third of its positions for civil servants following a large-scale action of that kind in Hesse. The classification procedure - applied by the office holders themselves - was a relatively analytic one with seven different criteria classes and various possibilities for variation. The staff council - the personnel representation board - also wanted to be heard in the assessment action. Result of the action was an upgrading of 2,500 positions and the downgrading of about 500. For Hamburg this meant an extra expenditure of 4 million DM in the personnel field just for 1968. The whole thing did not lead to the ex-

pected general satisfaction - quite the contrary was the case: Only one year later 3,500 remonstrations had been submitted by the respective office holders. Conditions in the state of Hesse, which had served as an example for Hamburg, had taken a similar course in the meantime; approximately 56 per cent of the reassessed positions had to be raised with a corresponding boom of costs. This, in turn, had the result that as early as January 1968 the Hessian decree on position classification had to be rescinded again.

Such improvements on state level or for various parts of the public service - as for instance for teachers - inevitably also gave rise to high hopes among servicemen toward the possibilities of analytic position assessment. A multitude of efforts by the Association ensued, among them resolutions of General Meetings and addresses to the Ministry of Defense and the ministry in charge of pay matters for the public service - the Ministry of the Interior.

In the meantime, both ministries had initiated activities of their own. In the Ministry of Defense, the Force Structure Commission on Manpower had taken up work - which already has been the subject of two presentations before the MTA. One year later the Study Commission on Reforms of Public Service Law began extensive investigations under the Minister of the Interior. Proposals resulting from this work to a large degree also aimed at an analytic position assessment and found a - though still somewhat vague - definition in an action program in 1976.

The general aims of this reform program are:

(1) Improving performance of the public service.

(2) Reduction of costs.

(3) Realization of the principle of function-oriented pay.

(4) Measures for the improvement of personnel management - especially in placement - and pay.

To this end, three intruments are to be developed and will be expected to attain the following partial objectives:

- Function-oriented position assessment.
- Development and establishment of a function-oriented training.
- Function- and performance-oriented efficiency rating and pay for the individual position holders.

The program report also stresses the point that development of instruments is difficult and can only be solved over a longer span of time; in the course of it, the Ministry could fall back on important earlier studies, however. The latter, inter alia, referred to investigations conducted by the armed forces.

In addition, a report by a position assessment commission called into being by the Minister of the Interior was presented the end of 1972. In this report, offices or positions in in a meaning relevant to pay legislation were defined as a "combination of a wide range of activities of uniform valence (value) levels". The sweeping term of "office" - the report goes on to say - does indeed make an assignment of individual duty positions difficult. Nevertheless, abolishment of existing deficiencies and realization of equal pay for duties of

like valence would remain the goal. To achieve this, the report proposes pay to be determined more than before by functions performed and therefore to define the requirements of the individual office or position more accurately. Function- or job-descriptions should serve this purpose. The report presented 27 examples.

I would like to show you two examples of the A3 pay level, which is Obergefreiter - Private First Class, and of the A7 level, which is the rank of Oberfeldwebel - Sergeant First Class. I shall later return to this matter and ask you to compare both examples with the Association's concept on this subject.

The report also includes a model for a uniform assessment with three "requirement groups" subdivided into seven "requirement classes", which again comprise ten "steps". Groups, classes and steps are accorded percentage values and also a step value code ranging from 0 to 300. Also, there is a point system with an attainable minimum of 80 points and a maximum of 1,000. The points reached by a position in assessment decide its assignment to a certain pay group.

This report, too, points in its summary to the diversity of difficulties and to the still pending assessment problems which cannot be solved in a short time. A, and I quote "determination of values which is strictly objective and free of human opinions and judgment", is deemed impossible. All assessment standards at best would be "decision aids for the legislator".

In this regard that many changes have taken place that, in an effort to realize the new legal principle of function-oriented pay, the Minister of the Interior appointed fifteen working groups to work out suggestions for function assessment. The military are not included in this program. This means that the danger of grave disadvantage again arises.

4 - 1976 Stock-Taking by the Deutscher Bundeswehr-Verband

At the end of 1975 the Association set up a committee on pay matters which I have the honor of presiding. Representatives of all three services and all groups of rank were appointed to this committee. Most of the committee members had already participated in five working groups, which for two days had investigated the present central pay problems affecting the armed forces. Under consideration of the working groups' findings and of 66 resolutions on pay matters passed by the last General Meeting of the Association, the committee was assigned the task of developing a long-term pay concept. A stock-taking at the very first session of the committee revealed that because of the many provable "pay inequities and absurdities" the main effort would have to be the development of principles for a well-balanced assessment of positions. Necessity for this is already established by the reasons cited in the new pay law for implementation of the new principle of function-oriented pay; they are inter alia:

(1) The present practice of applying the principle of combining wider fields of functions under one abstract title of an office with a uniform level of value variedly and dependent upon the respective employer (or part of the service) has partly led to diverging assessments of equal functions in

the past.

(2) The objective of achieving a more function-oriented pay system through establishment of a uniform (grade or promotion) limit of positions has not been reached.

(3) The present influence of educational and training background upon assessment must be reduced in accordance with function-oriented factors, which means replacement by the entire characteristics of the respective profession.

(4) Assignment of functions should be effected in several phases. To begin with, the most important typical functions of the various fields should be covered and considered.

The above paragraphs 3 and 4 were of special importance for the considerations of the committee regarding position assessment.

The committee therefore first took up examination of known methods for job or position classification in order to establish their usefulness for the analysis and assessment of duties and positions typical of the armed forces. During this process committee members examined more than a dozen methods, among them the "Navy's Point Method"; special attention was paid to the time-and-methods system (REFA) and to a plan forwarded by future general staff officers at the Bundeswehr Command and General Staff College years ago but never put to the test. The committee came to realize that none of the presently known methods were able to meet the requirements adequately. A number of tests with the time-and-methods system, for instance, showed that it can give useful results for the lieutenants' level and

maybe even for the position of a captain of the specialist officer category. All higher positions, though, and especially those with leadership tasks and responsibility will have to be handled decidedly different at least in grading and basic assessment. This would inevitably create the impression of having applied different methods and the charge of discrimination arising as a result would be difficult to clear away.

The committee put down its initial findings in eight short thesises:

(1) The principle of uniformity in public service law may not be departed from.

(2) The Association's concept must meet the demands of the servicemen adequately in a twofold way:
- Especially urgent problems must be solvable in a short time.
- The long-term concept has to be a system adaptable to changes necessitated by developments.

(3) The requirements of a position should not be measured against the imponderables of the state of defense - the war.

(4) The system of extra allowances should be simplified and standardized wherever possible by function-oriented classification and pay for peacetime position requirements.

(5) Examinations of certain typical positions are to provide a general perspective of the many special requirements peculiar to the armed forces. In case of certain keystone positions as that of the squad leader, so far unsolved classification problems for typical tasks in the forces must be analyzed in order to permit a function-oriented

classification within the framework of public service law. In the course of this no single branch should be favored or discriminated. Functions should always be viewed as part of the overall picture.

(6) Examination of requirements typical of the armed forces has to consider three major fields above all:
- The activity field of LEADERSHIP
- The activity field of ADMINISTRATION
- The activity field of TECHNOLOGY.

These fields often overlap in the armed forces. Thus, work and requirements of all three fields may apply to one position. This results in additional assessment difficulties.

(7) The professional profile of soldiering needs no closer examination, but special hardships typical of it should be carefully considered.

(8) Independent of the efficiency principle and function-oriented position classification, the principle of safeguarding the income standards of the individual should be observed as part of public service law because otherwise greater changes in classification will lead to strong opposition in the forces.

The committee experienced the latter already after a first sketchy analysis of the tasks of the squad leader and publication of a corresponding article in the Association magazine under the heading of "Is the Squad Leader Correctly Assessed?" and showing a chart titled "Requirements in Comparison - Pay Sergeant versus Squad Leader". In this article the

"profession" characteristics of the squad leader were first briefly listed: We have a total of about 24,000 squad leaders in the Bundeswehr, that is almost 5 per cent of all servicemen. Of these squad leaders, roughly two thirds have the rank of Unteroffizier - a junior NCO roughly equal to corporal (OR-5 lower half)-or Stabsunteroff zier - something like senior corporal or Airman First Class (OR-5 upper half) - if they serve longer, and only one third have the rank of Feldwebel - Staff Sergeant (OR-6). The latter are not primarily military leaders with leadership tasks, though, but rather with predominantly technical assignments, for instance squad leaders of maintenance and repair squads. The following thus does not apply to them but to the majority of squad leaders - approximately 17,000 - which have the following function: Lowest level of military command and control hierarchy, key position for realization of all command decisions which have to be carried out on a wide basis. Key position, because the squad leader is the only military superior in constant contact with the soldier at the base and able to exercise constant influence on his squad.

Short training and low pay for these men has caused press comments calling the Bundeswehr "technologically a giant, but pedagogically (or educationally) a dwarf". This impression stems from the fact that the bulk of the roughly 250,000 conscripts leaving the Bundeswehr annually has predominantly or exclusively come into closer contact with only these superior.

Looking for the squad leader with the highest requirement level, one automatically is confronted with the squad leader

of the armored infantry squad with the rank of Unteroffizier or Stabsunteroffizier, whose numerous tasks will be explained later. A graphic presentation of the demands put on such a squad leaders and using the demands on a pay sergeant as yard-stick of comparison and considering the requirement fields of ABILITY, RESPONSIBILITY and STRAIN shows - as you can see - the qualification requirements of the armored infantry squad leader with three exceptions overwhelmingly in front. The pay sergeant is only required to possess greater knowledge and ability in his non-technological specialty. He is required to act more independently in the field of responsibility in his specialty. And third, he is charged with higher and more direct responsibility in the field of accountability for action taken. But apart from these three exceptions the requirements for the function of the squad leader are considerably greater , without being duly considered in assessment of the position. This is evidence by the fact that the pay sergeant - other than the squad leader - is rated Staff Sergeant and Sergeant First Class and paid accordingly. Publication of these problems in the Association magazine led to a wave of protests from countless pay sergeants as well as from squad leaders in other arms or branches of service.

Because of this, already in the next issue of the magazine the special situation of the squad leader was explained in detail, especially in regard to the additional complications which had arisen for him since Basic Military Service had been lowered from 18 to 15 months in 1971: The aim then was to create more conscription equity without sacrificing combat power and

combat readiness of the armed forces. For this purpose the branches of service were directed to streamline and rationalize basic training. The burdens ensuing were hinted at in the 1973/74 White Paper on the Security of the Federal Republic of Germany and the Development of the Federal Armed Forces as follows: "The former 18-months replacement has been succeeded by a quarterly replacement cycle. As a result, post-basic training is no longer a continuous process but rather a sequence of recurring exercises whose object it is to integrate the newly arrived personnel in a combat team." But who has to shoulder this burden and what - in comparison - are its effects? Making this comparison of requirements between the squad at the time of the 18-months basic military service and that of the present 15-months term and again with the squad leader of the armored infantry as assessment base, the graph depicting the resultant changes clearly reveals the following additional complications:

- In present consolidated basic training, performance has to be double that of the earlier system.
- Time available for selection and training of the individual recruit for his special assignment in the combat team has shrunk to an extent where a mistake in selection or placement becomes almost irreparable.
- The post-basic training pahse - advanced special training and unit training - takes place under 12 months of unabating pressure for required readiness. This comes about because the squad theoretically does not change its general level of training and efficiency through the quarterly recurring changes in its personnel roster - in other words, it has

to keep up its readiness at a uniform level.

- The constant necessity of integrating 25 per cent newcomers in the combat team joined with the parallel burden of having to deal with 25 per cent of personnel to be discharged - this double and additional requirement constitutes a combination of psychological and educational demands completely new in the Bundeswehr.
- And finally, the educational master demand on the squad leader of having to train his squad during every post-basic quarter in various classes of different levels at the same time.

I may summarize this subject. Because of the new force structure - that is reduction of basic military service - the pass-through of conscripts has increased annually by about 55,000 recruits, under the following aggravating circumstances:

- Without an according increase of instructors,
- but by increasing qualitative training requirements at the same time through combining the former general basic training and special basic training to the new consolidated basic training.

This implies that within three months, that is half of the former training period, the same training objectives have to be achieved because otherwise the 12-months NATO-readiness obligation cannot be met. Since all of these things went hand in hand with a moving-up and shortening of NCO-training as well as lowering the minimum time for promotion to NCO, the result may be summed up as follows:

The Unteroffizier - the junior NCO - is either a miracle from wonderland with these characteristics: a very young squad

leader with less experience and shorter training who can train more soldiers better in shorter time; or he will remain the ideal of wishful thinking with a hoped for efficiency bitterly needed but mostly achieved in words only.

Squad Leader in Performance Comparison

Training - Experience - Length of Assignment - Efficiency

This chart shows that:

(1) Military leaders must be able to transform skills and knowledge gained in training into practical ability, that is efficience. Pressure-cooker methods in squad leader training do not suffice, they remain stopgaps.

(2) The ability to assume leadership responsibility includes certain personality factors which can be gained primarily through professional experience - that is length of assignment.

(3) The strain, especially physical stress, normally is mastered adequately also by the inexperienced and only shortly trained squad leader.

(4) Improvement of performance as envisioned by the action program of the Ministry of the Interior or the endeavoured higher combat power and combat readiness cannot be achieved with draftee-NCO's or short-term volunteers. Rather, a performance deficit through overtaxing the powers of many squad leaders must be suspected.

Another publication in the Association magazine stressed the importance of valuation:

Requirements in Comparison

Pay Sergeant versus Squad Leader

In this reversed think-model the squad leader and not the pay sergeant posed as measuring stick. Since the squad leader does not have to deal with any one of the demanding administrative tasks of the pay sergeant it is easy to discern that this time the picture is completely reversed. Solely in the requirement field of STRAIN the requirements are higher for the squad leader. Contrary to the model presented earlier - with the pay sergeant as yardstick, this time the requirement field of GENERAL PREREQUISITES (BASIC REQUIREMENTS) was added to the comparison. This is to consider the educational and training courses the pay sergeant has to run through prior to his assignment. They, too, show that more efforts are invested in the training of the pay sergeant. This striking disproportion, demonstrated by the two models, points to one of the central problems of position assessment, namely

Positions in Comparison

Leader versus Specialist

This comparison of two structurally totally different positions illustrates in abstract form one of the difficult problems. The military leader is something of an all-round man with a great number of diversified tasks of medium distinction or grading - shown in the chart between A2 and A8, whereas the specialist acts in a narrow, defined and related field of tasks with expert special knowledge. To find a well-balanced

and fairly just method of assessment for both parties poses a key problem of military position classification.

The Association next published a comparison of advantages and disadvantages of summary respectively analytical position assessment in order to inform its members about the problematics involved. The summary assessment was explained as follows: The degree of difficulty of an activity is appraised lump-sum. Advantages are simple handling, requiring relatively little time, primarily orienting itself by the training required - level and duration of it constituting the guiding principle of the assessment. Disadvantages, however, easily result from faulty evaluation since differing criteria for the difficulty of various tasks in one function may be misinterpreted or even overlooked. This method also widely precludes an objective comparison of assessments of dissimilar activities. What this can lead to in the armed forces has been clearly demonstrated with the charts "Requirements in Comparison - Squad Leader versus Pay Sergeant".

In the analytic assessment methods, in contrast to the above, requirements are examined separately and only after that the partial assessments are joined to a total assessment following a predetermined system. The most evident disadvantages are: Up to now there exists no method for a uniform analytic assessment in the public service of the Federal Republic of Germany. Development of such a method costs a lot of work, money, time. Realization and application next require patience, understanding, and again time. However, these disadvantages are

offset by decisive advantages:

- Uniform assessment criteria.
- Basis also for an objectified individual evaluation of performance.
- Possibility for constant refinement of the systematics.
- Repression of the imponderability factors job situation and political influences.
- Reduced danger of conflicts.
- Objectivation of differences of opinion.

As contrast to the description of a function provided by the Ministry of the Interior let us now turn to the "job" description of a squad leader of the armored infantry:

> "The function of the squad leader requires him to make and keep one tenth of the company in accordance with the pertinent regulations physically and psychically operational, combat ready and leadable".

What hides behind these words? For realization of this organizational function the position holds five major tasks.

(1) To lead the squad mounted or dismounted, with or without combat vehicle during exercises (or in the state of defense with such dexterity and mobility that it can be employed tactically useful in company action.

(2) To train the squad in accordance with regulations and manuals in such manner as to make it operational by ability and knowledge.

(3) To direct and supervise preventive maintenance of arms and equipment as well as technical care of the combat

vehicle in such manner that the squad is also optimally operational in arms and equipment.

(4) To supervise general company duties in such manner as to ensure team spirit of the squad.

(5) By leading through example in training, supervision and in exercises and thus exerting such influence on the squad that the soldiers find reaffirmation of the meaning of their military service and thus not only are operationally ready but also defense motivated in terms of the mission of the Bundeswehr.

A short analysis of the requirement field ABILITY with the skills, knowledge and aptitudes the squad leader needs, leads us to three basic requirements, namely:

(1) Qualification for the grade of NCO with corresponding personal, mental and physical characteristics.

(2) Completed formal education.

(3) Completed squad leader training and adequate experience.

He has to master or know well two dozen regulations and a dozen training manuals.

By the aid of twelve examples I would like to sketch the typical requirement characteristics:

(1) Mastery of the operational principles of the armored infantry platoon in all modes of combat.

(2) Knowledge of data and effectivity of own and hostile weapons.

(3) Mastery of combat training and part of weapons training materia.

(4) The ability to recognize essentials quickly and to transform them rationally into action, if necessary under time pressure.

(5) The ability to organize simple activity processes in such manner that personnel and equipment are employed functionally.

(6) Knowledge of the basic elements of modern leadership.

(7) Knowledge of the essential possibilities of care and welfare.

(8) Ability to promote human relations, to motivate subordinates to maximum performance and to ensure discipline and order in the squad.

(9) Skills and knowledge in all fields in which he is employed as instructor.

(10) Ability and experience in assessing performance, recognition of mistakes and in correcting them.

(11) Basic understanding for technological functions.

(12) Practical knowledge in all care and maintenance activities connected with combat vehicle, arms and equipment of the squad.

The scope of RESPONSIBILITY is outlined in the following requirements:

(1) Willingness to assume responsibility for ten soldiers under his command as well as for arms and equipment worth about DM 1 million (squad with combat vehicle MARDER) during training on post and during exercises.

(2) Knowledge of own powers of command and of the limits of obedience.

knowledge of all safety regulations of importance to life and health of soldiers assigned to him, as well as the ability to recognize dangers quickly and to react sensibly.

The main requirements in the field STRAIN require the following of the squad leader:

(1) The ability to assess changing situations correctly, make decisions and transform them into action so that missions or problems of the squad can be quickly accomplished or solved - all this under conditions of physical stress reaching the limit of his strength.

(2) The ability to act almost always under the pressure of lack of time, to transform orders of superiors into action and to enforce orders, to be constantly in the current of tensions between superiors and subordinates and yet not to lose his readiness for initiative.

(3) The readiness to work much longer hours than normal on many occasions and of having to perform unforeseen tasks in the course of this.

(4) The preparedness to subject himself and the soldiers serving under his command to numerous inconveniences - to dangers for health and life, to bear with extreme weather conditions and to accomplish missions without regard to noise, dirt, dust etc; to work for a sustained period of time in confined conditions (trench, combat vehicle), often in cramped position.

The activities of the squad leader outlined so far already give an impression of the special demands to be made on an

instrument for definition, analysis and assessment of military positions.

Catalogue of Characteristics

You see here a catalogue of characteristics, developed by the committee on pay matters. Its distinction is that it comprises less than 30 requirement classes. This makes it comparatively handy. Its relevance to the requirements encountered in the armed forces is underlined by the fact that the first requirement in the field of ABILITY is tactical leadership and action. For reasons of personnel management, as well as career and pay systematics especially the requirement classes

- educational background requirements
- training requirements
- requirements for experience
- special requirements

have been considered in the field of ABILITY.

Definition of the individual requirement classes and their grading or valuation are at the same time difficult and important. The Association has also made attempts to provide guidelines in this respect. For instance:

Administrative activity

You see here six steps or levels, beginning with "very low" and running up to "extremely high" with explanation of the various criteria. For instance at the level of "very low": Very few, very simple and repeating tasks without special earlier knowledge; can be performed after very short period of familiarization". In comparison, the definition for "extremely high": Very much above average, demanding and quite

unusual new tasks, which additionally to "very high" require further extensive and versatile experience as well as organizational ability, respectively innovational ability.

5 - Thesises on Position Assessment

The Federal Armed Forces Association and its committee on pay matters realize that the Association by itself cannot develop a perfect instrument for the assessment of positions in the armed forces quickly and unaided. Because of this the Association for the time being confines its efforts to submission of thesises on position assessment. These thesises contain the basic considerations to be observed. One of the essential points is the timely and encompassing inclusion of the Association in the development and especially testing and application of the instrument for position assessment. Tests conducted by the Association so far with a variety of methods have proved not entirely satisfactory; this only underlines the difficulty of the problem. Special difficulties, also perceived by the Association, exist in the necessity for long-term personnel planning and development in the forces and especially in the officer category, where positions with varying value have to be filled consecutively without disadvantages being permitted to arise for the individual. This means that position assessment in the forces must confine itself to a basic assessment in order to permit - if necessary - a pay open to rise and linked to rank, because otherwise immobility in the armed forces would reach an extent where they would become inoperational as far as personnel management is concerned.

In closing, permit me to touch briefly upon an important success of our Association in this field of activity. A publication of the Ministry of Defense toward the end of September of this year acknowledged for the first time that the squad leader - as the Association had pointed out repeatedly - is indeed underrated and should receive higher classification, even if the financial posture does not yet permit it at this time. The Association will submit comprehensive and workable suggestions for solution on this subject in the near future.

Gentlemen, I thank you for your attention.

OCCUPATIONAL ANALYSIS FOR NAVY INSTRUCTIONAL SYSTEMS DEVELOPMENT (ISD): A MATRIX APPROACH
(A Sound-on-Slide Presentation)

D. D. DAVIS, Staff, Naval Education and Training Command
T. M. ANSBRO, Staff, Naval Education and Training Program Development Center, Ellyson, Pensacola

OCCUPATIONAL ANALYSIS FOR NAVY INSTRUCTIONAL SYSTEMS DESIGN "ISD"

A MATRIX APPROACH

Since 1973, a major objective of the Chief of Naval Education and Training (CNET) has been: To provide a proficient occupant in terms of job/skill certification for every billet in the fleet.

TO PROVIDE A PROFICIENT OCCUPANT FOR EVERY BILLET IN THE FLEET

726

SLIDE #1
664

In order to meet this goal, the CNET realizes that the requirements of Navy jobs, as they are, and as they should be, must be determined.

OCCUPATIONAL ANALYSIS TO DETERMINE PRESENT AND FUTURE JOB REQUIREMENTS

727
SLIDE #2
665

As early as 1962, the Navy's traditional observation/interview/ technical conference approach to job analysis had become too expensive. The main drawback, however, was the fact that data obtained through these procedures was in non-quantifiable format and could not be analyzed satisfactorily.

SLIDE #3
666

In 1958, the United States Air Force made a most significant breakthrough in occupational research--the adaptation of computer technologies to the input, storage, and retrieval of job data. This program, known as the Comprehensive Occupational Data Analysis Program (CODAP), has continued to expand to this date.

1958

USAF COMPREHENSIVE OCCUPATIONAL DATA ANALYSIS PROGRAM

"CODAP"

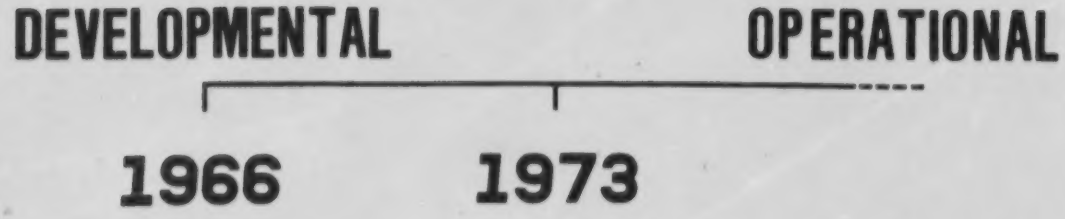
729

SLIDE #4
667

The Chief of Naval Operations (CNO) began laying the groundwork for a Navy Enlisted Occupational Analysis System and a Navy Occupational Data Bank in 1966. Beginning with the aviation ratings, the Navy Occupational Task Analysis Program (NOTAP) became fully operational in 1973. By early 1977, NOTAP will have collected, stored, and displayed occupational data for all Navy enlisted ratings. It will then begin its second iteration.

NAVY OCCUPATIONAL TASK ANALYSIS PROGRAM

"NOTAP"



730

SLIDE #5
668

The Naval Education and Training Command (NAVEDTRACOM) considers NOTAP data to be by far the single most comprehensive data base for use in the analysis phase of Instructional Systems Design (ISD).

SOURCES FOR NAVY ISD

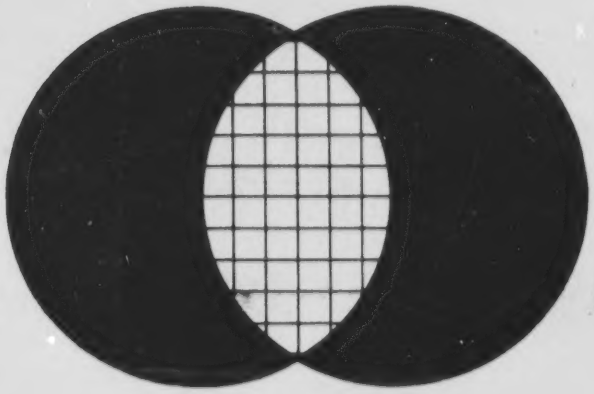
- NOTAP
- CONSOLIDATED SHIPBOARD ALLOWANCE LISTS
- HUMAN FACTOR STUDIES
- MANNING DOCUMENTS
- TECH MANUALS
- OCCUPATIONAL STANDARDS

For two reasons, the CNET requires that other viable data be used in conjunction with NOTAP computer printouts: (1) NOTAP data represents the "what is" in the Navy world of work. It was not designed to, nor could it, predict what the Navy world of work "should be." (2) It provides job/occupational information from representative samples of the Navy population, or by broadly defined ratings, not always by specific job assignments, which are inherently responsive to specific technological events.

NOTAP REPORTS THE "WHAT IS"

CNET CONSIDERS ALSO THE "WHAT SHOULD BE"

In order to develop a systematic approach to the use of all viable job data in the determination of job requirements within individual and among several related ratings, the CNET set aside, in August 1975, a group dedicated to the development of procedures for use in determining commonality and uniqueness of skills among all jobs in the Navy.

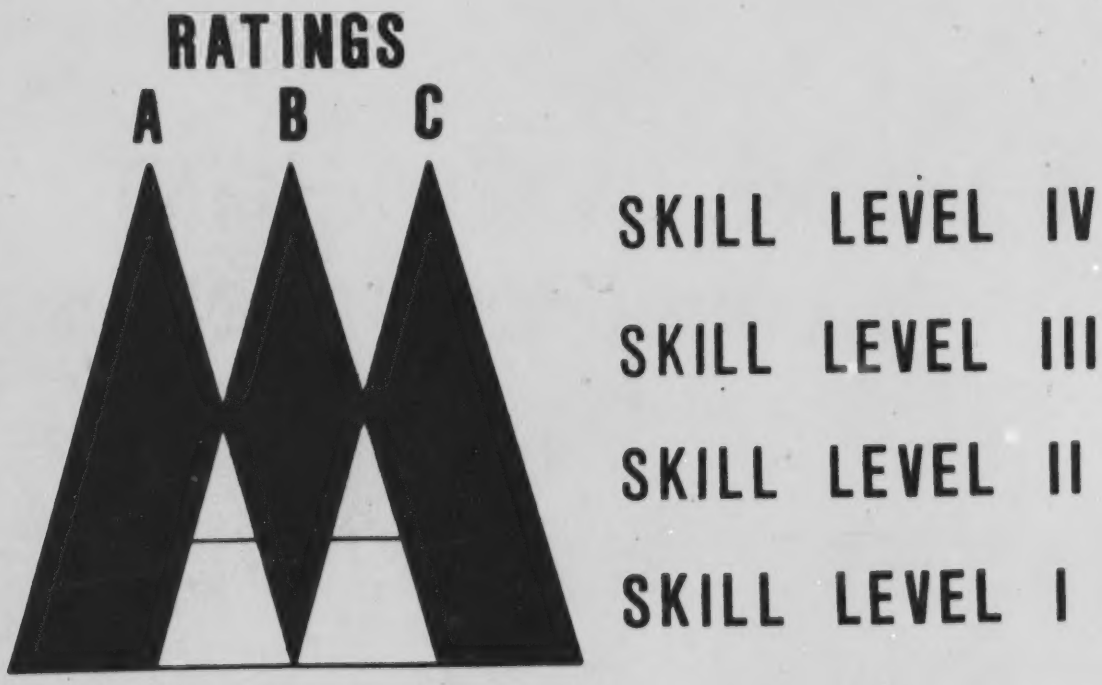


JOB SKILLS

733

SLIDE #8
671

This group, the Career Training Analysis Group (CTAG), has taken a rather innovative approach to job analysis in that it is striving to determine job skill requirements for apprentice, journeyman, supervisor, and management levels, for example, within and among ratings.



A more conventional approach would have been to analyze job requirements of predetermined groups within each rating in support of revising existing courses or of designing new ones. This approach, however, would NOT have provided a comparison of tasks and duties structured vertically within ratings and horizontally among, or across, ratings to support the design of career training paths.

RANDOM DEVELOPMENT OF INDIVIDUAL COURSES

OBJECTIVELY DERIVED TRAINING PATHS

As the initial project, the CTAG was assigned four of the avionics ratings. Early on, it was faced with the realization that data in the amounts necessary to describe skills required of incumbents in four ratings, from paygrades E1 through E9, could not be accommodated without computer assistance.

MULTIPLE-RATING ANALYSIS REQUIRED COMPUTER ASSISTANCE

736

SLIDE #11
674

This requirement drove the CTAG to devise a method of data tabulation which ensures uniform analysis by the generation of fixed categories of task statements and elements compatible with computer scanning requirements.

FIXED CATEGORIES OF TASK STATEMENTS:

JOBS
DUTIES
TASKS
SKILLS

737

SLIDE #12
675

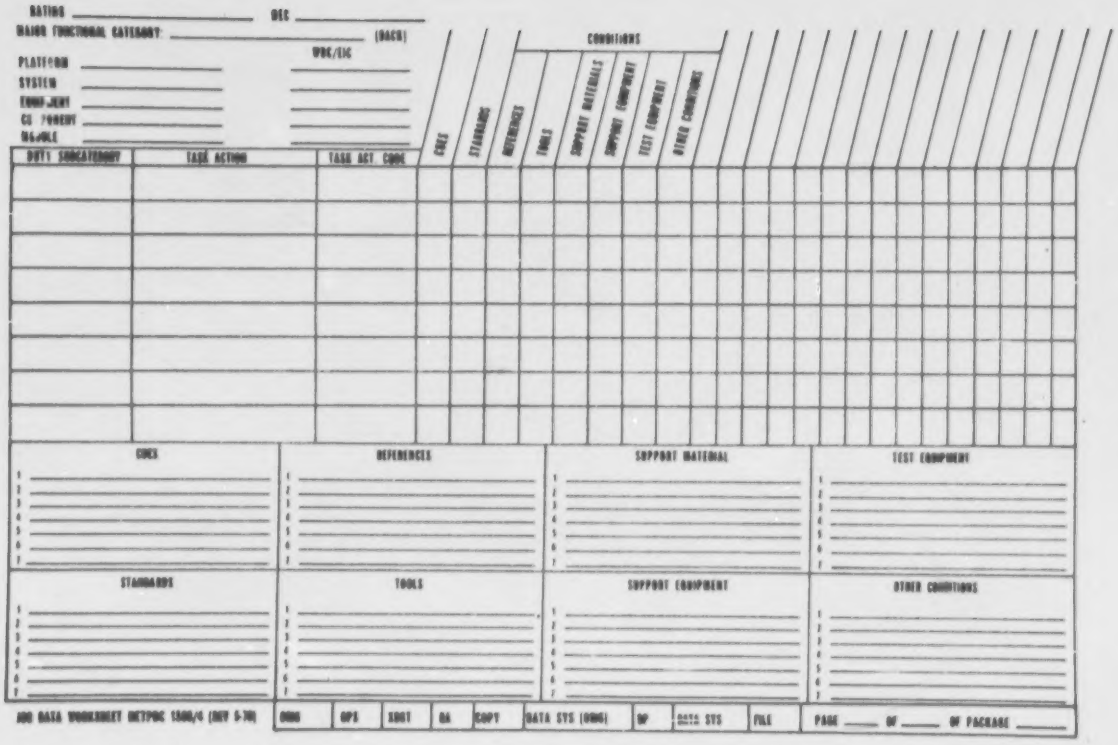
In December 1975, CNET obtained, through contract, services of the University of West Florida, to assist the CTAG in the development of procedures for shredding out job data for storage and subsequent retrieval through automated data processing. Dr. Richard EARP was assigned as the principal investigator.

JOB ANALYSIS DATA STORAGE AND RETRIEVAL THROUGH AUTOMATED DATA PROCESSING

738

SLIDE #13
676

You will be presented a "hands on" view of the matrix approach which is a direct result of the endeavors of Dr. EARP and the CTAG.



In the process of deriving fixed categories of work statements, the CTAG has identified five major categories under which all Navy duties can be placed. The major functional categories are: MAINTENANCE, FABRICATION, OPERATION, ADMINISTRATION, and MILITARY, each of which requires a specifically oriented matrix.

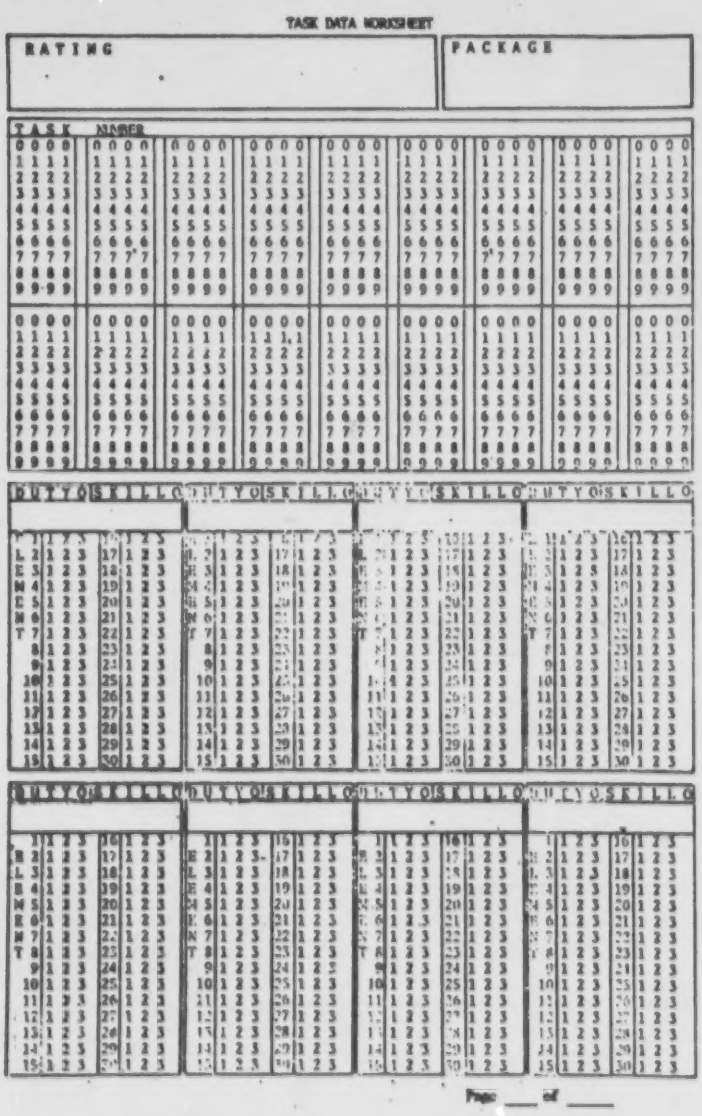
MAJOR FUNCTIONAL CATEGORIES

1. Maintenance
2. Fabrication
3. Operation
4. Administration
5. Military

740

SLIDE #15
678

There are two principal documents used in the CTAG approach to analysis: The Task Data Worksheet and the Job Data Worksheet. Shown here is the Task Data Worksheet, the analysis document of the two. On this document tasks are displayed with arrays of task data.



The second document, the Job Data Worksheet, is a bridging document which takes us from job documents to the Task Data Worksheet. On the Job Data Worksheet, several tasks may be entered, whereas a Task Data Worksheet is completed for each task.

RATING AT DOC AT-6671
BASIC FUNCTIONAL CATEGORY: MAINTENANCE (BASIC)
PLATFORM P-3B WUC/EIC APBC
SYSTEM H.F. COMMUNICATIONS 6/000
EQUIPMENT AN/ARC-94 6/180
COMPONENT
MODULE

| DUTY QUALIFICATION | TASK ACTION | TASK ACT. CODE | CREW | STANDARDS | REFERENCES | TOOLS | SUPPORT MATERIALS | SUPPORT EQUIPMENT | TEST EQUIPMENT | OTHER CONDITIONS | ACTUAL TIME (6/180) | MAINTENANCE MAN-HOURS (6/180) | MAN-HOURS (6/180) | MAN-HOURS (6/180) |

The upper left corner of the Job Data Worksheet is used for recording pertinent data, from the Navy rating involved, down to the specific element of concern--the actual equipment or component upon which actions are taken, or tasks performed.

RATING _____ NEC _____

MAJOR FUNCTIONAL CATEGORY: _____ (BACK)

	WUC/EIC
PLATFORM _____	_____
SYSTEM _____	_____
EQUIPMENT _____	_____
COMPONENT _____	_____
MODULE _____	_____

DUTY SUBCATEGORY	TASK ACTION	TASK ACT. CODE

In this example, the rating (AT) has been recorded, as has the platform (aircraft) on which the task actions are performed, and the Work Unit Code (WUC) or Equipment Information Code (EIC). Other codes, such as stock or supply codes, exist. As codes are merged, changed or updated, this information can be restated.

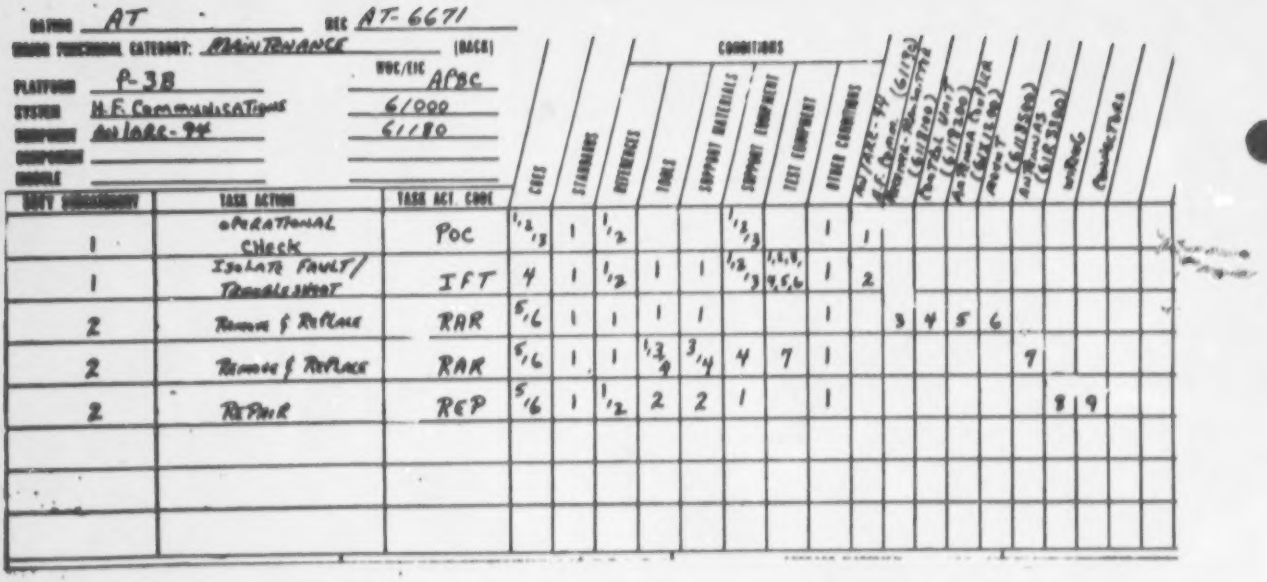
RATING AT NEC AT-6671

MAJOR FUNCTIONAL CATEGORY: MAINTENANCE (BACK)

		WUC/EIC
PLATFORM	P-3B	APBC
SYSTEM	H.F. COMMUNICATIONS	61000
EQUIPMENT	AN/ARC-94	61180
COMPONENT		
MODULE		

DUTY SUBCATEGORY	TASK ACTION	TASK ACT. CODE
1	OPERATIONAL CHECK	POC
1	ISOLATE FAULT/ TROUBLESHOOT	IFT
2	REMOVE & REPLACE	RAR
2	REMOVE & REPLACE	RAR
2	REPAIR	REP

The equipment being analyzed is the AN/ARC-94 communications set. This is entered in two places, as shown, on the Job Data Worksheet. In the remaining blank columns to the right of CONDITIONS are listed the components of the AN/ARC-94. These include the receiver-transmitter, antenna coupler, mount, and antennas.

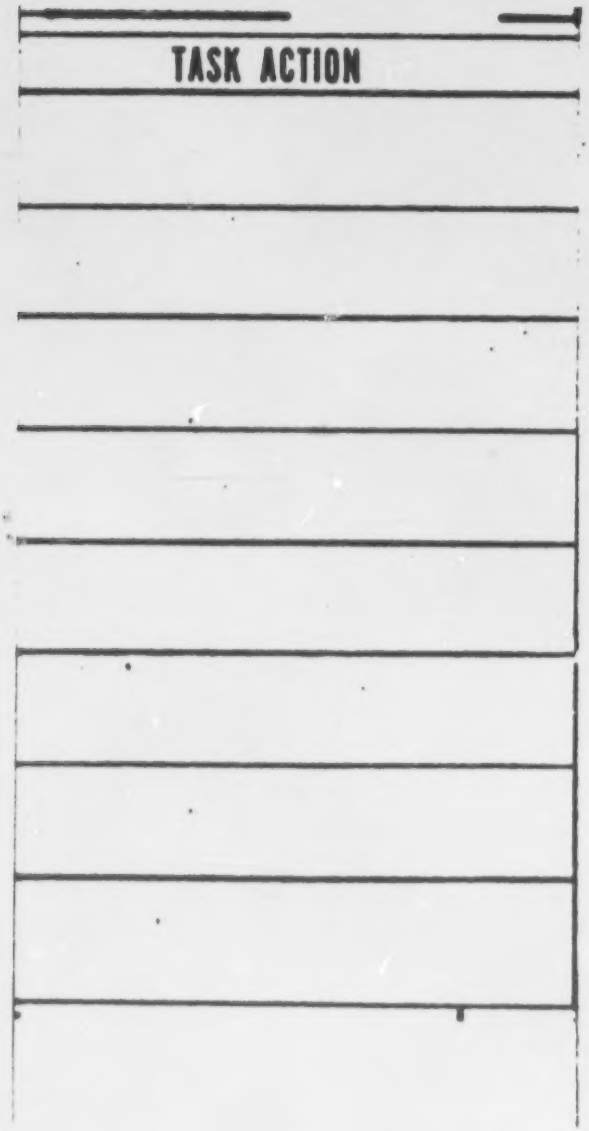


								CONDITIONS														
			CUES	STANDARDS	REFERENCES	TOOLS	SUPPORT MATERIALS	SUPPORT EQUIPMENT	TEST EQUIPMENT	OTHER CONDITIONS	SPARE #4 (6/178)	SPARE #5 (6/178)	SPARE #6 (6/178)	SPARE #7 (6/178)	SPARE #8 (6/178)	SPARE #9 (6/178)	SPARE #10 (6/178)	SPARE #11 (6/178)	SPARE #12 (6/178)	SPARE #13 (6/178)	SPARE #14 (6/178)	SPARE #15 (6/178)
SEQ	TASK ACTION	TASK ACT CODE																				
1	OPERATIONAL CHECK	POC	1,2,3	1	1,2			1,2,3		1	1											
1	ISOLATE FAULT/ TROUBLESHOOT	IFT	4	1	1,3	1	1	1,2,3	1,4,8, 9,5,6	1	2											
2	REMOVE & REPLACE	RAR	5,6	1	1	1	1			1		3	4	5	6							
2	REMOVE & REPLACE	RAR	5,6	1	1	1,3,9	3,4	4	7	1						7						
2	REPAIR	REP	5,6	1	1,2	2	2	1		1							8	9				

Associated or interrelated hardware, such as wiring, connectors, etc., authorized for repair at specified maintenance levels (organizational, intermediate, or depot) are entered also.

CONDITIONS																								
CDES	STANDARDS	REFERENCES	TOOLS	SUPPORT MATERIALS	SUPPORT EQUIPMENT	TEST EQUIPMENT	OTHER CONDITIONS	AD/AXC-94 A/E Comm. Monitor-Translator (6117B)	(6117B/100) Control Unit	(6117B/200) Antenna Coupler	(6XX1290) Mount	(6117B500) Antennas	(6117B300) Wiring	Connectors										
1,2,3	1	1/2			1/2/3		1	1																
4	1	1/2	1	1	1/2/3	1,2,3, 4,5,6	1	2																
5,6	1	1	1	1			1		3	4	5	6												
5,6	1	1	1/2/4	3/4	4	7	1						7											
5,6	1	1/2	2	2	1		1							8	9									

Reference publications contain step-by-step procedures or checklists which provide specific tasks to be performed on the equipment. These tasks are extracted and are recorded in the TASK ACTION column.



747

SLIDE #22

In this example, the tech manuals call for an operational check on the AN/ARC-94. That task "OPERATIONAL CHECK" is recorded. Use of the fixed category, TASK ACTION, on this matrix required that the CTAG derive, and use consistently, a definition of "task" as being composed of elements, each of which is a single, discreet behavioral action. Immediately to the left of the TASK ACTION is recorded a numeral that represents one of twenty duty subcategories which include: CHECKING/TESTING/INSPECTING, REPAIRING, CLERICAL, SUPERVISING/MANAGING, CONSTRUCTING, DESTRUCTING, STANDING WATCH, TRAINING, LOGISTICS, etc. The "1" shown here represents CHECKING/TESTING/INSPECTING.

DUTY SUBCATEGORY	TASK ACTION	TI
1	OPERATIONAL CHECK	
1	ISOLATE FAULT/ TROUBLESHOOT	
2	REMOVE & REPLACE	
2	REMOVE & REPLACE	
2	REPAIR	

Notice that "POC" has been recorded under TASK ACTION CODE. This is a CTAG code which represents Perform Action Check.

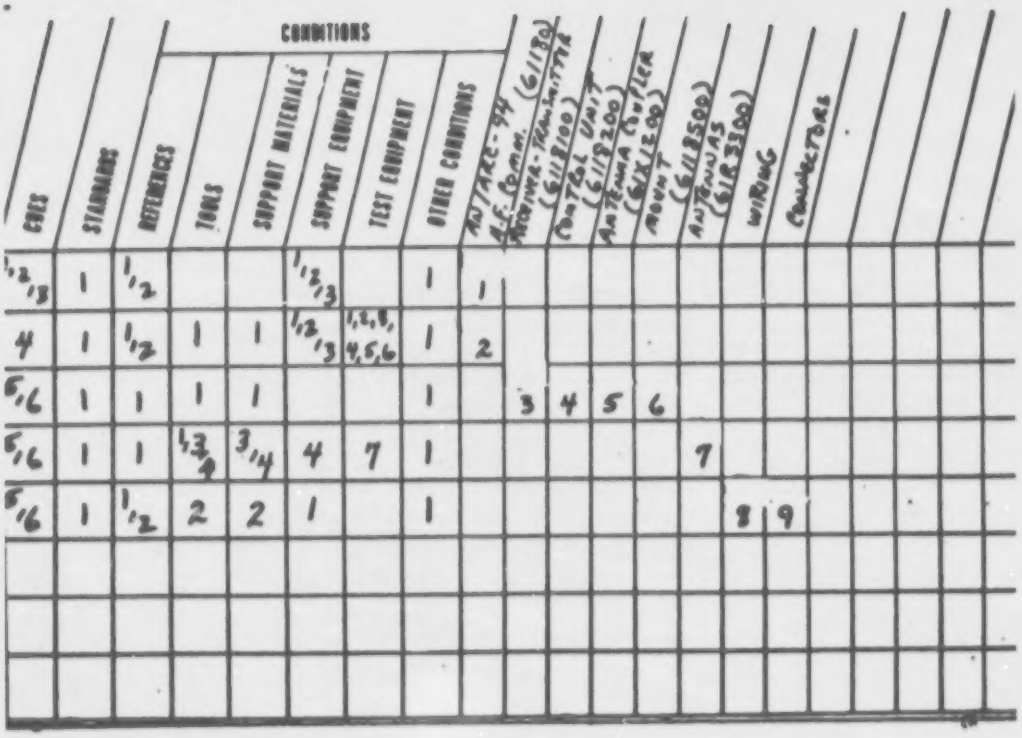
RATING AT NEC AT-6671

MAJOR FUNCTIONAL CATEGORY: MAINTENANCE (BACK)

		WUC/EIC
PLATFORM	P-3B	APBC
SYSTEM	H.F. COMMUNICATIONS	61000
EQUIPMENT	AN/ARC-94	61180
COMPONENT		
MODULE		

DUTY SUBCATEGORY	TASK ACTION	TASK ACT. CODE
1	OPERATIONAL CHECK	POC
1	ISOLATE FAULT/ TROUBLESHOOT	IFT
2	REMOVE & REPLACE	RAR
2	REMOVE & REPLACE	RAR
2	REPAIR	REP

The "1" under AN/ARC-94 tells us that the task indicated, "POC," is the first task performed in this sequence of tasks displayed on the Job Data Worksheet.



At the bottom of the Job Data Worksheet are blanks for entering data such as CUES, REFERENCES, SUPPORT MATERIALS, TEST EQUIPMENT, STANDARDS, TOOLS, and OTHER CONDITIONS. In this example, cues which prompt the performance of an operational check are "1, 2, and 3." These are recorded in the CUES column.

CUES

1. MALFUNCTION
2. MRC
3. REPAIR COMPLETE
4. OPERATIONAL CHECK
5. VISUAL INSPECTION
6. FAULT ISOLATION
7.

STANDARDS

1. I.A.W. REFERENCE PUBLICATION
2.
3.
4.
5.
6.
7.

JOB DATA WORKSHEET NETPDC 1500/4 (REV 5-76)

Similarly, numerals are recorded for each of the blocks. All these data have been extracted from official source documents, listed as REFERENCES.

CMS	REFERENCES	SUPPORT MATERIAL	TEST EQUIPMENT
1 MALFUNCTION	1 NAVAIR 01-75 PAA-2-6.1	1 SAFETY WIRE	1 DUMMY LOAD 82A
2 ADJ	2 NAVAIR 01-75 PAA-2-1	2 SOLDER	2 MULTIMETER AN/PSM-4A
3 REPAIR COMPLETE	3	3 SILICONE COMPOUND MIL-S-8660	3 SIGNAL GENERATOR AN/URM-89
4 OPERATIONAL CHECK	4	4 SEALING COMPOUND MIL-S-8516	4 TEST SET 678-1
5 VISUAL INSPECTION	5	5	5 VTVM TS-520/U
6 FAULT ISOLATION	6	6	6 WATTMETER AN/URM-120
7	7	7	7 TENSIOMETER (30 LBS)

STANDARDS	TOOLS	SUPPORT EQUIPMENT	OTHER CONDITIONS
1 T.O. 31R2-2ACS-1 REFERENCE PUBLICATION	1 COMMON TOOLS	1 POWER UNIT MC-1B	1 "O" LEVEL MAINTENANCE
2	2 SOLDERING IRON	2 AIR CONDITIONER MR-5	2
3	3 ANTENNA WIRE RETRIEVER # 19808	3 HEADSET/MIKE	3
4	4 MAST PLUG DRIVER # 19806	4 MAINTENANCE PLATFORM	4
5	5	5	5
6	6	6	6
7	7	7	7

NAVAIR DATA WORKSHEET METPOS 1300/4 (REV 5-70) | DOCS | OPS | LOGT | QA | COPY | DATA SYS (DOCS) | UP | DATA SYS | FILE | PAGE 1 OF 200 OF PACKAGE 0076

Figure 1

When a Job Data Worksheet has been completed for each of the tasks listed thereon, a Task Data Worksheet is completed for each task on the Job Worksheet which contains the specific duty subcategory and skills pertinent to the task being analyzed. More than one task can be recorded on a Task Data Worksheet only if the characteristics and behavior actions for both are identical. Information on the Task Data Worksheet is coded to a Master Reference Sheet which delineates specific duty subcategories and skills pertinent to task performance.

RATING A7 REC A7-6671

BASIC FUNCTIONAL CATEGORY: MAINTENANCE (BASIC)

PLATFORM P-3B WUC/EIC APBC

SYSTEM H.F. COMMUNICATIONS 6/000

EQUIPMENT AN/ARC-94 6/180

COMPONENT

GROUPS

DUTY SUBCATEGORY	TASK ACTION	TASK ACT. CODE	CUES	STANDARDS	REFERENCES	CONDITIONS: TOOLS	CONDITIONS: SUPPORT MATERIALS	CONDITIONS: SUPPORT EQUIPMENT	CONDITIONS: TEST EQUIPMENT	CONDITIONS: OTHER CONDITIONS	AUTOMATIC GO (6/1150)	MANUAL GO (6/1150)	AUTOMATIC NO GO (6/1150)	MANUAL NO GO (6/1150)	AMPLIFY (6/1150)	ANTENNAS (6/1330)	WIRING	PANEL/PLATES
1	OPERATIONAL CHECK	POC	1,2,3	1	1/2			1/2/3		1	1							
1	ISOLATE FAULT/ TROUBLESHOOT	IFT	4	1	1/2	1	1	1/2/3	1,4,7, 9,5,6	1	2							
2	REMOVE & REPLACE	RAR	5,6	1	1	1	1			1		3	4	5	6			
2	REMOVE & REPLACE	RAR	5,6	1	1	1/3/9	3/4	4	7	1						7		
2	REPAIR	REP	5/6	1	1/2	2	2	1		1							8	9

CUES
1. MALFUNCTION
2. ARC
3. REPAIR COMPLETE
4. OPERATIONAL CHECK
5. VISUAL INSPECTION
6. FAULT ISOLATION
7.

REFERENCES
1. NAVAIR 01-75 PAA-2-6.1
2. NAVAIR 01-75 PAA-2-1
3.
4.
5.
6.
7.

SUPPORT MATERIAL
1. SAFETY WIRE
2. SOLDER
3. SILICONE COMPOUND MIL-S-8660
4. SEALING COMPOUND MIL-S-8516
5.
6.
7.

TEST EQUIPMENT
1. DUMMY LOAD AN/UPM-4A
2. MULTIMETER AN/PSM-4A
3. SIGNAL GENERATOR AN/USM-44C
4. TEST SET 678-1
5. VTVM TS-510/U
6. WATTMETER AN/URM-120
7. TENSIOMETER (200 LBS)

STANDARDS
1. NAVAIR MAINTENANCE PUBLICATION
2.
3.
4.
5.
6.
7.

TOOLS
1. COMMON TOOLS
2. SOLDERING IRON
3. ANTENNA WIRE PULLER # 19206
4. MAST PULL DRAWER # 19206
5.
6.
7.

SUPPORT EQUIPMENT
1. POWER UNIT MC-1B
2. AIR CONDITIONER AR-5
3. HEADSET / MIKE
4. MAINTENANCE PLATFORM
5.
6.
7.

OTHER CONDITIONS
1. "O" LEVEL MAINTENANCE
2.
3.
4.
5.
6.
7.

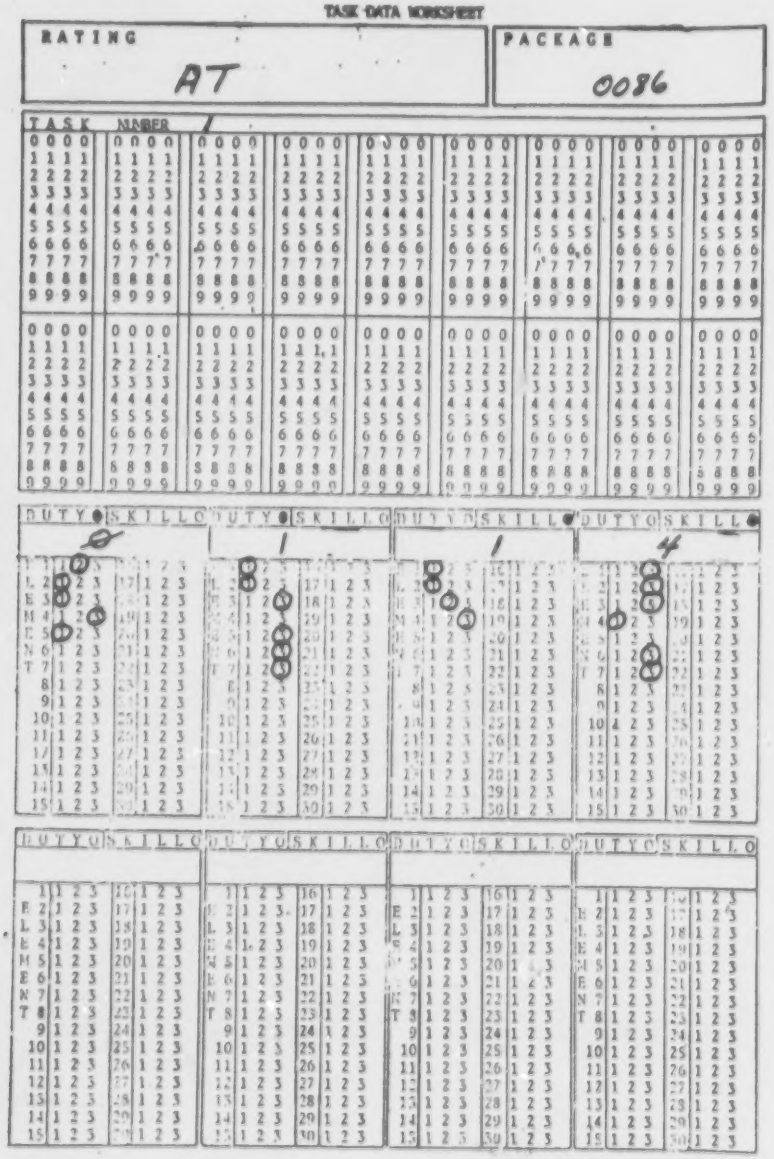
DD DATA REQUIREMENT DEFINE 1000/4 (DEF 5-73) | BASIC | OPS | XMIT | QA | COPY | DATA SYS (CODE) | MP | DATA SYS | FILE | PAGE 1 OF 200 OF PACKAGE 0076

For ref 1

Twenty "Duty Subcategories" and five "Skill Areas" have been listed on the back of each Job Data Worksheet. Each of these can be entered on the Task Data Worksheet.

DUTY SUBCATEGORIES			SKILLS
General	7. Constructing	14. Performing honors/ceremonies	1. Using references
1. Checking/testing/inspecting	8. Destructing	15. Standing watches	2. Using tools
2. Repairing	9. Counseling	16. Performing first aid health functions	3. Using support materials
3. Operating equipment for designed mission	10. Training	17. Performing disaster preparedness functions	4. Using support equipment
4. Performing clerical functions	11. Performing safety functions	18. Performing damage control functions	5. Using test equipment
5. Supervising/managing	12. Performing security functions	19. Housekeeping (general)	
6. Designing	13. Performing logistic functions	20. Performing other assigned functions	

For purpose of illustration, you will be shown a Task Data Worksheet which has been used for recording supportive data for the task, "Perform an operational check on the AN/ARC-94." We will go through the essential steps required for entering the data.


TASK DATA WORKSHEET

RATING: AT

PACKAGE: 0086

TASK NUMBER: 1

DUTY / SKILL

SLIDE #30

The task number "1" indicates that this is the first task performed.

TASK DATA WORKSHEET

RATING	PACKAGE
AT	0086

TASK NUMBER

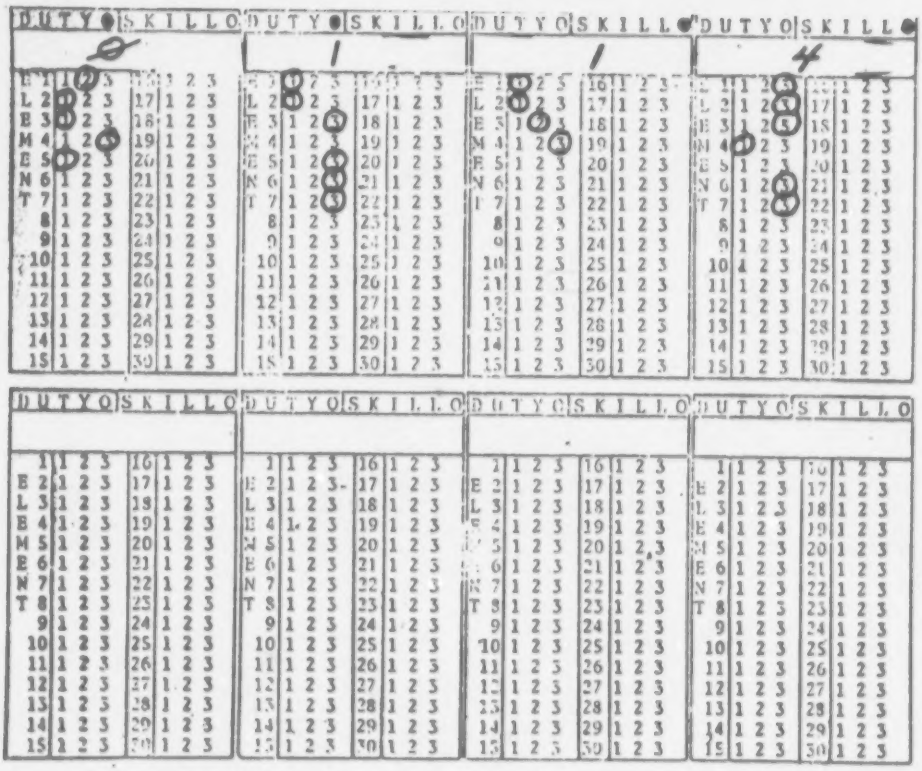
0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9
0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8 9 9 9 9

75C

SLIDE #31

694

The remaining columns are used for recording information supportive of duty subcategories and skill areas.



The "General" category (zero) is used for all tasks. Use of this category is indicated by use of a zero at the top of the first column. Since General is a Duty Subcategory, the circle opposite DUTY is blackened.

DUTY	SKILLO															
E 1	1 2 3	16	1 2 3		1 2 3	16	1 2 3		1 2 3	16	1 2 3		1 2 3	16	1 2 3	
L 2	1 2 3	17	1 2 3		1 2 3	17	1 2 3		1 2 3	17	1 2 3		1 2 3	17	1 2 3	
E 3	1 2 3	18	1 2 3		1 2 3	18	1 2 3		1 2 3	18	1 2 3		1 2 3	18	1 2 3	
M 4	1 2 3	19	1 2 3		1 2 3	19	1 2 3		1 2 3	19	1 2 3		1 2 3	19	1 2 3	
E 5	1 2 3	20	1 2 3		1 2 3	20	1 2 3		1 2 3	20	1 2 3		1 2 3	20	1 2 3	
N 6	1 2 3	21	1 2 3		1 2 3	21	1 2 3		1 2 3	21	1 2 3		1 2 3	21	1 2 3	
T 7	1 2 3	22	1 2 3		1 2 3	22	1 2 3		1 2 3	22	1 2 3		1 2 3	22	1 2 3	
8	1 2 3	23	1 2 3		1 2 3	23	1 2 3		1 2 3	23	1 2 3		1 2 3	23	1 2 3	
9	1 2 3	24	1 2 3		1 2 3	24	1 2 3		1 2 3	24	1 2 3		1 2 3	24	1 2 3	
10	1 2 3	25	1 2 3		1 2 3	25	1 2 3		1 2 3	25	1 2 3		1 2 3	25	1 2 3	
11	1 2 3	26	1 2 3		1 2 3	26	1 2 3		1 2 3	26	1 2 3		1 2 3	26	1 2 3	
12	1 2 3	27	1 2 3		1 2 3	27	1 2 3		1 2 3	27	1 2 3		1 2 3	27	1 2 3	
13	1 2 3	28	1 2 3		1 2 3	28	1 2 3		1 2 3	28	1 2 3		1 2 3	28	1 2 3	
14	1 2 3	29	1 2 3		1 2 3	29	1 2 3		1 2 3	29	1 2 3		1 2 3	29	1 2 3	
15	1 2 3	30	1 2 3		1 2 3	30	1 2 3		1 2 3	30	1 2 3		1 2 3	30	1 2 3	

DUTY	SKILLO															
	1 2 3	16	1 2 3		1 2 3	16	1 2 3		1 2 3	16	1 2 3		1 2 3	16	1 2 3	
E 2	1 2 3	17	1 2 3		1 2 3	17	1 2 3		1 2 3	17	1 2 3		1 2 3	17	1 2 3	
L 3	1 2 3	18	1 2 3		1 2 3	18	1 2 3		1 2 3	18	1 2 3		1 2 3	18	1 2 3	
E 4	1 2 3	19	1 2 3		1 2 3	19	1 2 3		1 2 3	19	1 2 3		1 2 3	19	1 2 3	
M 5	1 2 3	20	1 2 3		1 2 3	20	1 2 3		1 2 3	20	1 2 3		1 2 3	20	1 2 3	
E 6	1 2 3	21	1 2 3		1 2 3	21	1 2 3		1 2 3	21	1 2 3		1 2 3	21	1 2 3	
N 7	1 2 3	22	1 2 3		1 2 3	22	1 2 3		1 2 3	22	1 2 3		1 2 3	22	1 2 3	
T 8	1 2 3	23	1 2 3		1 2 3	23	1 2 3		1 2 3	23	1 2 3		1 2 3	23	1 2 3	
9	1 2 3	24	1 2 3		1 2 3	24	1 2 3		1 2 3	24	1 2 3		1 2 3	24	1 2 3	
10	1 2 3	25	1 2 3		1 2 3	25	1 2 3		1 2 3	25	1 2 3		1 2 3	25	1 2 3	
11	1 2 3	26	1 2 3		1 2 3	26	1 2 3		1 2 3	26	1 2 3		1 2 3	26	1 2 3	
12	1 2 3	27	1 2 3		1 2 3	27	1 2 3		1 2 3	27	1 2 3		1 2 3	27	1 2 3	
13	1 2 3	28	1 2 3		1 2 3	28	1 2 3		1 2 3	28	1 2 3		1 2 3	28	1 2 3	
14	1 2 3	29	1 2 3		1 2 3	29	1 2 3		1 2 3	29	1 2 3		1 2 3	29	1 2 3	
15	1 2 3	30	1 2 3		1 2 3	30	1 2 3		1 2 3	30	1 2 3		1 2 3	30	1 2 3	

Page ___ of _____

Looking at the General Duty Subcategory Master Reference Sheet, we review the listed references to determine which of the areas apply. We determine that 1 through 5, SAFETY, SPEED OF RESPONSE, ACCESSIBILITY, TOLERANCE, and CONTINUITY OF ACTION apply to this task, as they apply to most tasks. Other areas such as SIZE/SHAPE (HANDLING) and WEIGHT are not considerations in the performance of this task.

(0) GENERAL*

1. SAFETY**

1. Violation should not result in damage to equipment or injury to personnel or safety is no factor.
2. Violation may result in the damage to equipment or injury to personnel.
3. Violation may result in both damage to equipment and injury to personnel.

2. SPEED OF RESPONSE**

1. Immediate response not required -- allows ample time for research if necessary; e.g., filling out a form, measuring distance to stationary target.
2. Quick response -- does not allow time for research. Action must be taken automatically from learned responses; e.g., playing a musical instrument, receiving a CW message, tracking a moving target on radar.
3. Instant response -- action must be taken as fast as humanly possible; usually an emergency condition; e.g., applying brakes to avoid an accident, ditching an aircraft.

3. ACCESSIBILITY**

1. Easily accessible; no factor in complexity of task.
2. Moderately accessible -- requires opening drawers, removal of plates, panels, boots, covers, or minor components.
3. Difficult to gain access -- requires disassembly.

4. TOLERANCE**

1. Not applicable -- no established tolerances.
2. Noncritical -- small error will not materially affect the performance. Tolerances generally established but not published.
3. Precise -- within allowable error (published tolerances).

5. CONTINUITY OF ACTION**

1. Action may start and stop without affecting overall performance; e.g., troubleshooting, posting data, changing a tire.
2. Action may start and stop but may reduce the overall performance; e.g., aligning, painting.
3. Once action begins, it must be carried to completion; e.g., landing aircraft, survival.

NOTE:
* General applies to all tasks
** Mark for all tasks.

(0) GENERAL (Continued)

6. MINIATURIZATION

1. Size is no factor -- item can be viewed with normal vision and manipulated without special tools.
2. Miniature -- item can be viewed with normal vision but requires special tools for manipulation.
3. Micro-miniature -- item requires magnifying device to be viewed and requires special tools for manipulation.

7. SIZE/SHAPE (HANDLING)

1. Size/shape is no factor in positioning.
2. Size/shape is important -- item can be maneuvered into position in only limited number of ways.
3. Size/shape is critical -- item can be maneuvered into position in only one way.

8. WEIGHT (HANDLING)

1. Light weight -- easily positioned with one hand, or no handling required.
2. Medium weight -- requires two hands to position.
3. Heavy weight -- requires more than one person or mechanical assistance to position.

9. FRAGILITY (HANDLING)

1. Infrangible -- not easily damaged; e.g., paperwork, tire, resistor.
2. Moderately fragile -- must be handled carefully to avoid damage; e.g., electron tube, diode.
3. Fragile -- must use extreme caution to avoid damage; e.g., integrated circuit, stratified semiconductors known as chips, CRT, jet eng.

10. COMPUTATIONS

1. Simple arithmetic computations involving application of independent functions; e.g., add, multiply, subtract, and divide.
2. Complex formulas involving application of multiple arithmetic functions; e.g., $V = \frac{4}{3} \pi r^3$
3. Problem solving involving determination and application of a complete formula; e.g., determine rate of change of distance between two moving objects.

FIG. 2

Under the first area, SAFETY, we determine that "2", Violation of Safety Precautions, "may result in the damage to equipment or injury to personnel." This is recorded by circling the "2" to the right of the "1" which represents SAFETY. The same procedure is followed in the recording of other appropriate data under SPEED OF RESPONSE, ACCESSIBILITY, etc.

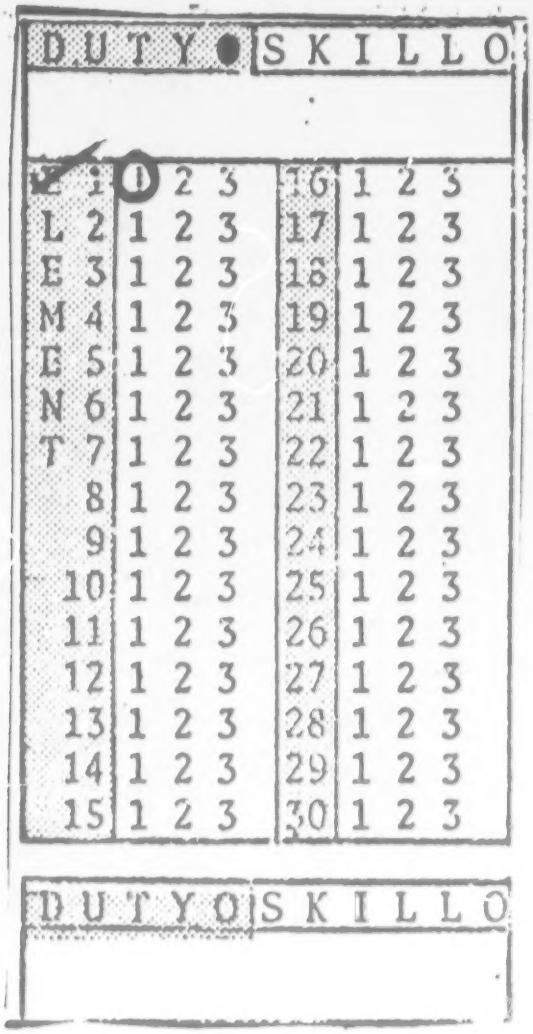
(O) GENERAL*

1. **SAFETY****

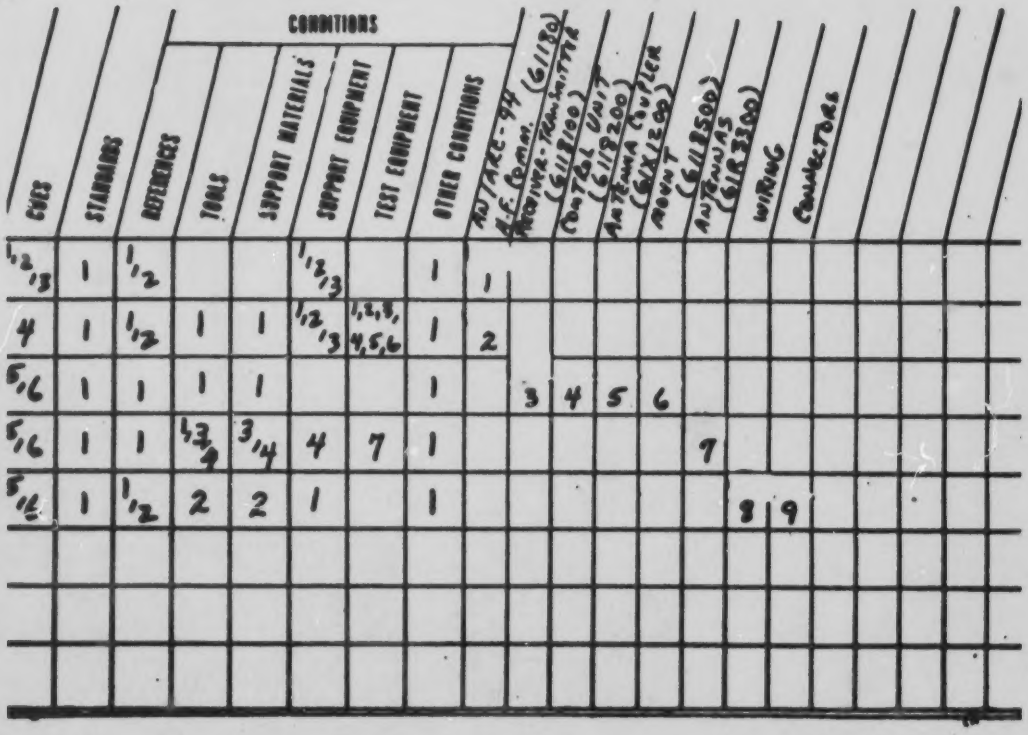
   1. Violation should not result in damage to equipment or injury to personnel or safety is no factor.
   2. Violation may result in the damage to equipment or injury to personnel.
   3. Violation may result in both damage to equipment and injury to personnel.

After GENERAL information has been recorded, we go directly to the Job Data Worksheet to determine which Duty Subcategory has been entered for the task. Duty Subcategory "1" has been entered. We blacken the circle opposite DUTY and review the Master Reference Sheet for Duty Subcategory "1" CHECKING/TESTING/INSPECTING. Through review of the references, we determine that "1, 2, 3, 5, 6, and 7" apply to the task; now, we circle the appropriate descriptor for 1, 2, 3, 5, 6, and 7 to further explain each area which applies to the task.

For 1, the first descriptor applies--"Analyze static indications only." We circle "1" on the Task Data Worksheet. The same procedure is followed in recording data for the other areas. We study references to provide specific information; we locate that information on the Master Reference Sheet; and we indicate on the Task Data Worksheet in the specific data which applies.



After having entered all relevant information for Duty Subcategory 1, we look back at the Job Data Worksheet under CONDITIONS at the top of the sheet. Each of the first five headings, REFERENCES through and including TEST EQUIPMENT, which have numerals under them cues us to the Master Reference Sheet for that category. For example, under REFERENCES, we find "1" and "2"; this indicates that two references listed at the bottom of the Job Data Worksheet are being used.



We then go to the Master Reference Sheet headed "SKILL AREA (1): USING REFERENCES." A careful review of this sheet and the reference material aids us in determining which of the skills listed apply to the task being analyzed. For each area which applies, we specifically determine its direct application and then record our determination on the Task Data Worksheet, just as we did under the Duty Subcategory Areas.

SKILL AREA: (1) USING REFERENCES

1. **SELECT/LOCATE REFERENCE(S)****

   1. Identification provided and reference(s) readily accessible in work area/technical library designated.
   2. Identification provided, but reference(s) must be acquired from external sources; or
      Requires research to make selection, but reference(s) readily accessible in work area/technical library designated.
   3. Requires research to make selection, and reference(s) must be acquired from external sources.

2. **LOCATE DATA (WITHIN REFERENCE)****

   1. Information is indexed and sequentially arranged within one reference.
   2. Information is indexed but branches throughout the reference(s); or,
      Information is not indexed but is sequentially arranged in one reference.
   3. Information is not indexed and branches throughout multiple references.

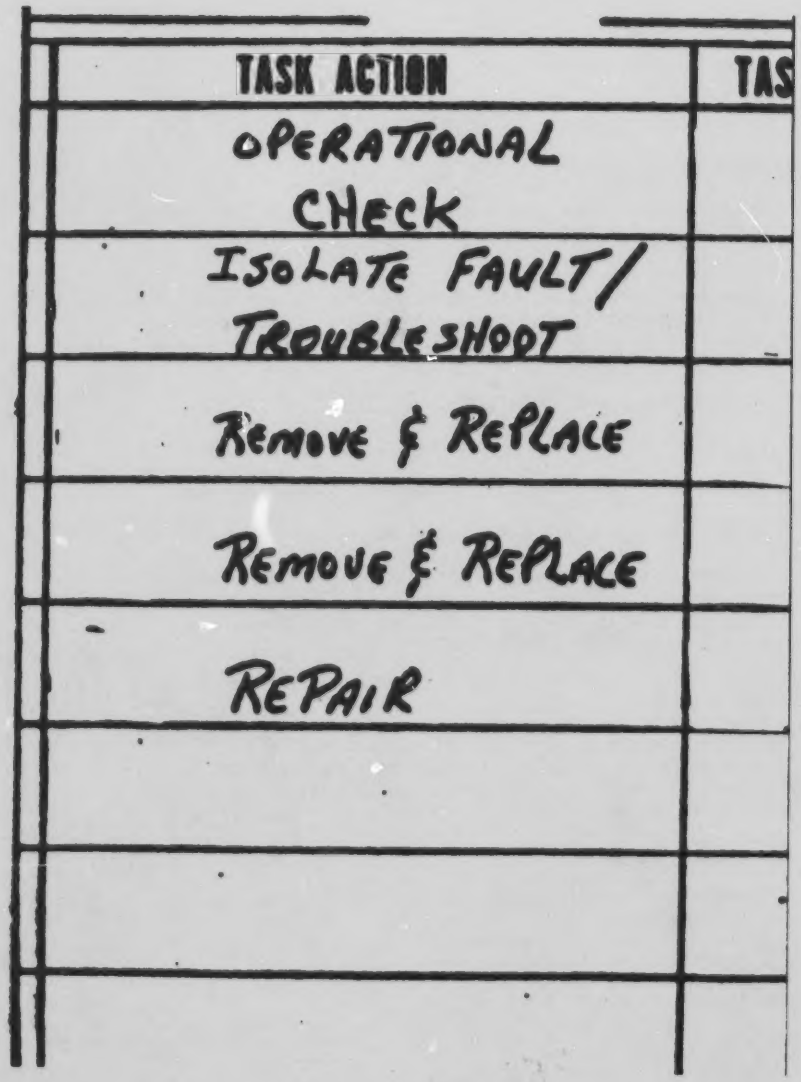
3. **DETERMINE FACTS/PROCEDURES****

   1. Locate facts/figures specifically stated; e.g., tolerances, safety factors.
   2. Locate specifically stated procedural information; e.g., checklist or operational check, MRC cards.
   3. Determine procedural information which is not specifically stated; e.g., determine troubleshooting procedures for an item for which no step by step procedures provided.

4. **INTERPRET LISTS/TABLES/DIAGRAMS**

   1. Interpret lists/tables composed of narration/numerals; e.g., checklists.
   2. Interpret lists/tables composed of symbols; e.g., bar graph, periodic tables of elements; or
      Interpret diagrams composed of narration/numerals; e.g., organizational chart.

We have gone through the most essential steps which result in utilization of a Job Data Worksheet matrix and a Job Task Data (OPSCAN) Worksheet for the recording of job data. For each task entered on a Job Data Worksheet, a Task Data Worksheet is completed to include relevant Duty Subcategories and Skills.



Each Job Data Worksheet and Task Data Worksheet is reviewed by a quality control (QC) team for accuracy and content. Once given an upcheck by the QC team, worksheets are delivered, presently, to the University of West Florida, where the job data are stored in the Florida State University System Computer. Various printouts are ordered from the University by the CTAG.

SLIDE #41

Where will this analysis and data storage procedure take the Naval Education and Training Command? It will provide a completely auditable display of job data, with all necessary elements subject to retrieval. It is hoped that massive areas of commonality will be identified in duty and task areas which require the same of similar skills. It will confirm or dispel the suspicions on the part of some that commonalities abound at the entry and apprentice levels and that complexity increases with the development of higher level skills. For training, it may provide objective insight into learning difficulty. At least, Navy will have a bank of tasks and skills required in the performance of job assignment, which should, as a minimum, support the development of job-relevant training which places skills and job requirements in harmonious balance, certifying the decision to accept or to reject existing hierarchies within the Navy's world of work.

JOB KNOWLEDGE ANALYSIS: A NEW APPROACH

Joseph S. Tartell, Captain, USAF
Hendrick W. Ruck
Walter E. Driskill, Ph.D.

JOB KNOWLEDGE ANALYSIS: A NEW APPROACH

Background & Purpose

The conventional approach to job analysis is initiated with the development of a comprehensive list of tasks. The analysis of job incumbents responses to this list of tasks provides the data required to determine the jobs that personnel perform. An alternative approach may be the use of an analysis of the knowledge required to accomplish jobs.

In order to evaluate this alternative approach, a universal model for evaluating the electronics fundamental training provided to airmen in the United States Air Force was developed by the USAF Occupational Measurement Center[1]. The model is expressed in terms of what knowledge is required for airmen to perform tasks in their assigned occupational specialty.

The first operational application of the Electronics Principles Job Inventory (EPI) was to the Space Systems Command and Control (AFS 308X0) career ladder. This occupational specialty was chosen for five reasons:

1. The specialty is small and has job incumbents performing a broad spectrum of tasks requiring varied amounts of electronics knowledge,
2. Most incumbents are stationed in the continental U.S.,
3. The management of personnel in the career ladder is in the process of being changed from command management to AF-wide management,
4. There is a requirement to develop a basic resident technical training course to include some electronics principles training, and
5. A recent occupational survey of the career ladder detailing its structure and specific tasks performed is available.

The purpose of this paper is to describe and report the ability of the Electronics Principles Job Inventory to aid in the determination of the structure of an occupational specialty. That is, through the use of an instrument which measures the type of knowledge required to perform tasks, is it possible to determine the functional structure of that occupational specialty?

Procedures

The Electronics Principles Job Inventory (EPI) was administered to 174 airmen in the Space Systems Command and Control career ladder during December 1975. Personnel were instructed to answer the items in the inventory with regard to the knowledge they use or refer to in their present job.

The inventories were administered in two ways: (1) Administration to personnel at Lowry AFB, Buckley ANGB, Offutt AFB, and Loring AFB was conducted in person by staff members of the Occupational Survey Branch; and (2) Administration to personnel at other locations was accomplished through procedures normally utilized for administration of job inventories by the Occupational Measurement Center, which involve mailing the inventories to the base personnel offices and having the Occupational Survey Control Officer administering the inventories to job incumbents.

Data analysis was accomplished through the use of the Comprehensive Occupational Data Analysis Programs (CODAP). In addition, the job analysis data available from the Occupational Survey Report (OSR) of the Space Systems Command and Control Career Ladder[2] completed in September 1975 were used.

Two specific areas of the data are the subject of this paper. First, did the functional job structure of the electronics principles survey match the functional job structure of the task analysis survey? And second, do the same people appear in the same jobs in both analyses?

Data Presentation

I. Analysis of the Electronic Principles Job Inventory responses yielded the following job clusters and job types:

Operations (36% of the sample, cluster)

- Superintendents and MAJCOM Supervisors (7%)
- Antenna Tracking Operators (3%)
- Ground Satellite Monitors (18%)
- Programmer Operators (4%)
- Data Analysts (4%)

Maintenance (48% of sample, cluster)

- Maintenance Technicians & Instructors (22%)
- Maintenance Technicians (17%)
- Instructors (5%)
- Maintenance Supervisors (4%)

Quality Control & Instruction (10% of sample, cluster)

- Quality Assurance Monitors (3%)
- Operations Instructors (7%)

A small number of cases (12) representing 6% of the sample were not included in the cluster analysis. These cases represented those respondents whose jobs were distinctly unique and could not be determined based on knowledge analysis.

II. Analysis of the Occupational Survey Report (OSR) data indicated the following job clusters and job types:

Maintenance (36% of sample, cluster)

- DSP Maintenance Technicians (16%)
- SAC Site Maintenance Technicians (8%)
- DSP Maintenance Specialists (7%)
- SAC Group Maintenance Technicians (5%)

Supervision (21% of sample, cluster)

- Superintendents & NCOIC I (18%)
- Superintendents & NCOIC II (3%)

Operations and Instruction (44% of sample, cluster)

- Data Analysts (8%)
- SAC DMSP Instructors (3%)
- Ground Satellite Monitors (13%)
- Primary Read-In Operators (7%)
- Programmer Operators (8%)
- Antenna Tracking Operators (5%)

Realigning the job clusters clearly demonstrates that the same jobs were found in each analysis.

EPI Analysis	Job Analysis
Operations (42%)	Operations/Instruction (44%)
Maintenance (48%)	Maintenance (36%)
QC and Instruction (10%)	Supervision (21%)
Not identifiable (6%)	

Further realignment of the job types indicates that the EPI is a useful tool for identifying specific functions within this career ladder.

EPI Analysis	Job Analysis
Antenna Tracking Operators (3%)	Antenna Tracking Operators (5%)
Ground Satellite Monitors (18%)	Ground Satellite Monitors (13%)
Programmer Operators (3%)	Programmer Operators (8%)
Data Analysts (4%)	Data Analysts (8%)
Superintendents & MAJCOM NCOICs (7%)	Superintendents & NCOICs I (18%)
Maintenance Superivsors (4%)	Superintendents & NCOICs II (3%)
Quality Assurance Monitors (3%)	
Instructors (5%)	SAC DMSP Instructors (3%)
Operations Instructors (7%)	
Maintenance Technicians & Instructors (22%)	DSP Maintenance Technicians (16%)
	DSP Maintenance Specialists (7%)
Maintenance Technicians (17%)	SAC Site Maintenance Technicians (8%)
	SAC Group Maintenance Technicians (5%)

Another step in determining functional job structure was a comparison of the job incumbents, by name, in each sample. In comparing the EPI data to the job analysis clusters, the overlap of personnel was substantial.

JOB ANALYSIS CLUSTERS	N	NAMES IN BOTH SAMPLES	NAMES MATCH SAME JOB CLUSTER
Maintenance	70	45 (64%)	42 (93%)
Supervision	35	13 (37%)	12 (92%)
Operations	90	61 (68%)	45 (74%)
	195	119	99

A reverse comparison, the OSR data to the EPI clusters yielded much the same information.

EPI CLUSTERS	N	NAMES IN BOTH SAMPLES	NAMES MATCH SAME JOB CLUSTER
Maintenance	83	65 (80%)	43 (68%)
QA & Instruction	19	13 (68%)	9 (69%)
Operations	73	41 (58%)	35 (86%)
	175	119	87

In an attempt to further refine the comparisons between the EPI and job analysis samples, the name lists were compared by job type group. The job analysis job types indicated substantial overlap when compared to the EPI job types.

JOB ANALYSIS	N	NAMES IN BOTH SAMPLES	NAMES MATCH IN SAME JOB GROUP
DSP Maintenance Tech	27	12	11 (92%)
SAC Site Maintenance Tech	14	13	12 (92%)
DSP Maintenance Spec	11	7	7 (100%)
SAC Group Maintenance Tech	8	6	6 (100%)
NCOIC & Superintedent I	30	13	13 (100%)
NCOIC & Superintendents II	5	1	1 (100%)
Data Analysts	12	9	6 (67%)
SAC DMSP Instructors	5	5	5 (100%)
Group Satellite Monitors	22	12	12 (100%)
Primary Read-In Operators	10	4	4 (100%)
Programmer Operators	12	9	6 (67%)
Antenna Tracking Operators	9	4	4 (100%)

To complete the by name comparison, the job groups from the EPI data were compared to the name lists for the job analysis data. The overlap, again, was substantial.

EPI ANALYSIS	N	NAMES IN BOTH SAMPLES	NAMES MATCH SAME JOB GROUPS
Superintendents & MAJCOM Supervisors	13	7	3 (43%)
Antenna Tracking Operators	5	4	2 (50%)
Ground Satellite Monitors	31	21	12 (57%)
Programmer Operators	6	2	0
Maintenance Technicians & Instructors	38	29	21 (72%)
Maintenance Technicians	30	25	21 (84%)
Instructors	8	6	2 (33%)
Maintenance Supervisors	7	6	4 (67%)
Quality Assurance	5	--	--
Operations Instructors	12	--	--
Data Analysts	7	6	1 (17%)

Analysis and Conclusions

Through the use of chi-square tests, comparisons were made between the job incumbents identified as being in a given cluster of jobs (maintenance, supervision, operation) based on job analysis data compared to the cluster identified through the use of the Electronics Principles (EPI) data. For these comparisons, the job analysis was considered "test 1" and the EPI was considered "test 2". There were 119 job incumbents who completed both "tests."

Significant relationships were found between the task-based job inventory and the electronics principles inventory.

Overall, the job cluster a given individual would be included in based on an analysis of his required job knowledge can be predicted with 62 percent accuracy (X^2=166.14, ϕ=.68, p<.001). Similarly prediction of which task-based cluster an individual is in based on job knowledge requirements may be made with 76 percent accuracy for maintenance personnel (X^2=45.54, ϕ=.62, p<.001), 62 percent accuracy for operations (X^2=53.11, ϕ=.67, p<.001), and 44 percent accuracy for supervisors (X^2=70.57, ϕ=.77, p<.001). It is important to note that the prediction of tasks groups individuals belong to based on job knowledge is significantly better than predictions based on probability alone. That is, the prediction accuracies of 62 percent (overall), 76 percent (maintenance) and 62 percent (operations), and 44 percent (supervisors) are considerably better than those of chance, 33 percent (overall), 53 percent (maintenance), 36 percent (operations), and 11 percent (supervisors). These results are surprisingly good when one considers the many factors serving to attenuate the relationships. For example, since the inventories were administered eight months apart, individuals may have changed jobs in that time.

Simply stated, one can predict accurately the job an individual performs given that one knows the knowleges that individual uses on his or her job.

These results indicate that for this electronic specialty there is a very strong relationship between the ability of a task analysis instrument to discern job groupings and a knowledge analysis instrument to discern the same type of job groupings. In addition to the ability of these two instruments to produce highly similar job groups, prediction

with reasonable accuracy can be made as to the job group into which an individual will fit using either instrument. Therefore, at least in this area of electronics knowledge, the use of either instrument will provide a valid assessment of the work individuals accomplish in the completion of their jobs and the training necessary for personnel to be able to perform these jobs.

REFERENCE NOTES

1. O'Connor, et al., A Universial Model For Evaluating Basic Electronic Courses In Terms Of Field Utilization Of Training, USAF Occupational Measurement Center, Lackland AFB, TX.

2. Keeth, J., Occupational Survey Report Of The Space Systems Command And Centrol Career Ladder, USAF Occupational Measurement Center, Lackland AFB, TX, 1975.

JOB KNOWLEDGE TESTING AS A MEASURE OF TRAINING EFFECTIVENESS[1]

EARL S. STEIN[2]

U.S. Army Natick Research and Development Command

Natick, MA 01760

[1]Presented at the 18th Conference of the Military Testing Association Gulf Shore, Alabama; October, 1976

[2]Coworkers on this project have included Dr. Lawrence Symington, Dr. Herbert Meiselman, Dr. Thomas Nichols and Mr. Ted Mattas.

JOB KNOWLEDGE TESTING AS A MEASURE OF TRAINING EFFECTIVENESS

Introduction

Does training of personnel meet the goals set by the organization that has to pay for it? This is frequently asked question when the time arrives to evaluate training programs. In many cases, however, the question may be close to unanswerable unless a system had been devised as part of an overall plan to examine the results of training. Do training graduates meet the performance criteria of the jobs they are being trained for? Often a more apt question is: Are there any performance criteria? The need for objective training criteria is especially relevant to organizations in which previous training evaluation was limited to managerial opinion or what in the military is often called command judgement. A systematic approach will at least not be inferior to the older intuitive methods (Tiffin, 1952), and may be employed as a source of additional information when managers are emotionally tied to judgemental training evaluation.

Criteria development is an age-old problem. Systematic training evaluations have employed attitude/opinion surveys, performance rating, work sampling, and measures of work output to mention a few. Thorndike (1949) and Foley (1974) cited three types of criteria: immediate measures, such as achievement tests; intermediate measures collected after training but not employing actual job tasks; and, finally, ultimate criteria or measures of real job performance. The distinction between these criteria are in many cases more a matter of degree of behavioral

specificity and timing relative to training than a function of psychometric excellence. Only the ultimate criteria, however, have the face validity of involving behavior samples of the job itself. Control of measurement on the job site can sacrifice some accuracy for relevance but the acceptable amount of this sacrifice is one that each investigator must decide upon when electing which level of criteria to employ.

Intermediate criteria have some advantages. The training evaluator does not have to work in the job setting itself, which means he has more control. The criteria are collected after training is completed thus allowing for normal forgetting which may not have had time to occur while the student was still in school. The expense of administration may be better controlled by having more freedom to choose when and where administration is to be done. The disadvantages of intermediate measures are also apparent. Actual job tasks are not employed, and this could reduce the validity of the measure. There are those who support the contention that only ultimate criteria have potential validity (Foley, 1974).

Job knowledge testing was selected as a technique for this current project. As an intermediate criterion, it allowed for potential sampling of many areas within the jobs specified in this study - Navy Food Service. We accept the fact that such testing is not an ultimate criteria and is at best an approximation to actual job performance. Its use is based on the premise of Gagne and Fleishman (1959) that underlying skills and knowledge give the individual the capability to do his job. As the tasks in a given work role increase in number and heterogeneity, it becomes

very difficult to deal with ultimate criteria (Vroom, 1964).

This study was concerned with the evaluation of the training received by Navy Food Service personnel. An effort was made to construct a Job Knowledge Test which reflected the variety and complexity of food service skills. Since many workers in this area may perform only a limited set of tasks out of the entire job spectrum, it was hoped that this test would sample a broader array of knowledge than could be gleaned by direct observation for a limited time. In theory and ideally a Navy Food Service worker is supposed to be a generalist capable of performing across all subskills in his job; but in practice he often does not get the opportunity.

Method

Test Construction.

The development of test items was considered a major stumbling block in the construction of the Job Knowledge Test (JKT). We had no subject matter experts that could be called upon to write items. An ideal source of such items would have been the Navy's Promotion Test system, which maintained a large test item pool with relevant test performance statistics. The controllers of this information declined to provide it on the grounds of test security. We were, however, offered the correspondence course materials normally employed by personnel to prepare for promotion exams. Although those materials were likely to be already in the fleet, it was decided to employ them based on several assumptions. First, sampled individuals would not know that job knowledge was to be tested until they were actually presented with the JKT;

and secondly, there was a large item pool from which to choose, thus decreasing the probability of recognition of a specific item. The correspondence course program provided no reliable item statistics. In order to select items for a pilot test, we had to read all the correspondence manuals and choose items on the basis of face validity and content coverage (Bureau of Naval Personnel, 1971a, 1971b; Naval Training Command, 1972, 1973). Only items relating directly to galley and wardroom performance were chosen. One hundred and thirty-nine items were initially selected. Slide 1 presents a typical item.

Slide 1

Pilot Test.

The first instrument was tested at a small naval base on the East Coast. Twenty-four participants who had considerable Navy food service experience responded to the test. Item difficulties were latter computed and items .80 and below .20 were eliminated in order to move the mean difficulty towards .50 which in theory provides the best inter-individual discriminations (Guilford, 1954, p. 365). This left 112 items in the test which were organized so that content was approximately balanced across the entire test.

Performance Scale Development.

In order to acquire some sort of validity estimate, it was decided to construct performance scales to be filled out by the supervisors of the personnel tested. We did not have the expertise to do a task analysis of food service jobs and elected to obtain the involvement of experts from the Navy Food Service System. Specifically, members of the Food Management teams on the East and West Coasts were viewed as having the

necessary skills to identify critical areas of behavior. The use of supervisors and internal experts is well known in industry and carries with it the added advantage of organizational involvement in scale development. The obvious result of employing such internal consultants is the improved face validity of the scales and the increased belief by organizational members that the scales are more that the product of ivory tower research that dictates arbitrarily to the consumer of psycho-technical services how to measure job proficiency (Barrett, 1966; Smith & Kendall, 1963; Williams & Seiler, 1973).

Originally, a three-stage program was envisioned to generate and develop adequate performance oriented scales. These stages included the development of scale categories or dimensions, the description of behaviors under each category, and the location of the scale value of the anchors using the Thurstone technique (Guilford, 1954). However time constraints on the initiation and completion of the study were severe and a decision was made to develop scale categories and use a general Likert seven point scale to assess each category.

A questionnaire was constructed entitled the "Navy Food Service Training Study: Critical Categories of Food Service Behavior". The instructions to the subject matter experts explained the importance of performance evaluation for training improvement in food service skills. These instructions noted that the NOTAP project (Navy Occupational Task Analysis Program) had provided extensive lists of relevant job tasks but that the terms (i.e., prepares salad dressing, cuts dough) were descriptive rather than evaluative and we were seeking classes of tasks that would lend themselves to observation and subsequent evaluation.

Those selected to participate as subject matter experts were asked to list up to twenty categories of behavior which met these criteria. They were encouraged not to make the list at one sitting, but to work on it when they had time. The importance of their contribution to the project was emphasized. Responses were collected from several food management teams and also from school instructors in Navy Food Service.

These responses were edited, and six categories were chosen based on the frequency of citation and also on their relevance for potential validation of the job knowledge scale. These included cooking ability, supervisory skills, motivation, personal hygiene, job knowledge, and attitude toward other workers. On each scale supervisors were asked to rate their personnel individually on a seven-point Likert scale. The scale was anchored as demonstrated in Figure 1.

Figure 1 about here

The use of such innocuous anchors as "good" and "bad" has often been controversial. Kavanaugh (1971) has pointed out that "practical limitations of time and money prevent the scientist/practitioner from following the ideal position of complete certainty of choice of stimuli based on thorough empirical information (p. 664)". The use of generic scale anchors was deemed necessary based on time limitations. Also, it has been demonstrated that raters can deal effectively with common terms such as "good" or "bad" although perhaps not as well with terms used for the middle of the scale such as average (Jones, Perrum, and Thurstone, 1955).

Subjects.

All participants were male enlisted men in pay grades E-1 through E-6. Three hundred and fifty-five food service specialists were sampled on the Atlantic and Pacific fleets. All personnel were informed of the purpose of the study and of their rights under informed consent doctrine. In general, personnel were very cooperative and attempted to complete the JKT. However, the time for administration had to be limited to one-hour because of practical restrictions on the degree to which we could interfere with normal shipboard operations. Radloff and Helmreich (1969) have noted that field data collection differs in many ways from the laboratory, and generally researchers have to work with operational units and avoid generating time constraints which commanders feel are abusive.

Administration.

The JKT was administered in common areas on ships such as mess decks (dining facilities) and in classrooms on shore. The place of administration was at the discretion of the appropriate commanders. The classroom facilities were generally adequate, but the conditions aboard ships varied considerably. Regardless of conditions, personnel seemed adapted to the environment and simply kept working.

Rating Scale Administration.

Supervisors were asked to write in the names of those of their men who were being tested that day. They were assured that the rating information would never be communicated back to the Navy personnel system by name and that as soon as statistical analysis were completed all names would be eliminated from our files. Being credable to the supervisors on these points was considered extremely important if our scales were to

make any unique contribution to the study over that already available from the Navy's enlisted rating system.

Results

Speeding and Intelligence Components.

During field administration of the JKT, time was limited to one hour. This limitation meant that at least for some of the personnel involved the JKT was a speeded test, because not everyone finished. Most completed between 80 and 100 of the 112-item instrument. There was some concern that the test was loaded with intelligence components as traditionally measured by paper and pencil instruments. This would have detracted somewhat from the value of the JKT as a job specific instrument and as an intermediate criterion of training effectiveness.

Scores on the Navy General Classification battery were available to us. This was a well standardized battery and no doubt measures some components of intelligence. Correlations between battery scores and those of the JKT were minimal as can be seen in Slide 2.

Slide 2

Reliability.

Alternate form and test-retest estimates of reliability were not practical possibilities. Therefore reliability was assessed using the Kuder-Richardson technique. The K-R reliability for the first 80 items, which virtually everyone completed, was .85. Since the instrument was made up of fairly homogeneous items, it was decided to estimate the reliability of the entire test using the Spearman-Brown formula. This

yielded a coefficient of .89. In order to verify the homogeneity assumption, individual scores were computed on the first 80 items and correlated against the total test scores. This coefficient was .91.

Item Analysis.

An item analysis was conducted on the first 80 items and it was assumed that the statistics were representative of the entire test. About 95% of the respondents completed these items. Data on 355 respondents were analyzed using the BMD factor analysis program (using orthogonal analysis with verimax rotation). The average item difficulty was .45 which was close to the desired .50.

The factor analysis generated ten rotated factors which could account for only 20% of the total variability. Many items were contributing to the total score in their own unique way and their contribution was minimally redundant with other items. This may have indicated that the sampled items covered a broad spectrum of the knowledge area.

Test Validation.

Test validation demonstrates a classic problem in applied industrial psychology, the selection of what Thorndike (1949) and Foley (1974) have called the ultimate criterion. Validity computations were made against two criteria neither of which could be perceived as really ultimate: the Navy efficiency report scores and our own behavioral rating scales. Slide 3 shows the relationships of the efficiency report scores to JKT test performance. These relationships were generally low and non-significant, except for three scales: ratings on instructor potential,

Slide 3

career counseling and independent duty. In contrast all the coefficients for our own scales were significant, although not great in magnitude. It was interesting at the anecdotal level that the highest correlation

Slide 4

was on the scale labelled "Job Knowledge". However, this may have involved some criterion contamination, because supervisors knew we were administering a test of job knowledge.

The rating scale data were factor analyzed using the same BMD program previously cited (Slide 5). One rotated factor was generated

Slide 5

which accounts for 70% of the variance in the ratings. Thus it was clear that supervisors were orienting on a single attribute of job performance as they reviewed the past successes and failures of their men. They did seem to discriminate between their subordinates, even if on only one attribute.

Pay Grade Effects.

The variability of average scores between pay grades or ranks within the sample was the clearest finding in this study. Slide 6 presents the

Slide 6

average scores for the entire sample of 355. There was a linear progression of scores from the lower to the higher pay grades. An analysis of variance of variance across pay grades showed a significant effect, $F(4,347) = 43.10$, $p \leq .01$, with a strength of association of $Omega^2 = .32$. This was a strong effect. The Newman-Keuls procedure was employed

to determine where the differences occured (Weiner, 1962). As can be seen in Slide 7, three groupings occured: E-2,3; E-4,5; and E-6. This

Slide 7

pay-grade effect was very consistent no matter how subgroups were taken out of the entire sample. There was some slight variability in the clustering of differences, but the ordinal relationship of pay grades remained consistent.

Educational Background.

Since the educational background of Navy Food Service personnel varied widely, the experiences of our participants reflected this variability. Education of these men in food service ranged from no formal instruction (primarily on-the-job-training) to correspondence courses and highly organized schooling. Comparisons were made between those who graduated from the basic school for food service personnel (known as "A" or Apprentice School) and those who had no formal training. Slide 8 shows that for each pay grade group listed, those who had been to A school produced higher JKT scores than those who had not received formal training. The exception to this was pay grade E-6 where no

Slide 8

significant difference was found. Also at the E-4 level the strength of association was not extremely high, although the difference was significant.

A second analysis was done on education variables dividing the sample into those who had: 1) no training, 2) correspondence training only, 3) formal school training only and, 4) both formal school and

correspondence training. With the pay grades pooled an ANOVA produced an F (3,352) = 14.98, $p < .01$ and an Omega^2 = .105. Post hoc testing was done using the Newman-Keuls procedure. Slide 9 presents the results. Regardless of what type of training experience an individual had, any

Slide 9

training was linked to better JKT performance than no training (including on-the-job training).

Discussion

The Job Knowledge Test was admittedly an impromtu instrument designed more quickly than any marketable standardized test. The test was assembled as a research tool and only as a first step in a larger training evaluation program. As an intermediate criterion, it suffered from the drawbacks of any measure not employing actual work performance. The next step would be the continued development of ultimate criteria involving behaviorally anchored rating scales and trained observer/raters. Also, other evaluative techniques such as measures of work output and product quality could be employed.

The effort to validate the JKT has had some limited success. The internal consistancy reliability was not perfect; but had it been, one might have wondered whether the instrument was too homogeneous relative to a job that contained many varied subtasks. Fortunately, the JKT was not significantly related to the Navy entrance battery which lent some hope to the possibility that it was not heavily loaded with IQ components. The criteria employed in our validation effort were at best of the concurrent variety. Our own scales were employed by supervisors on a memory basis rather than using direct observation of performance. This made them

similar administratively to the way traditional efficiency reports are completed. However, they were more current than the efficiency report data available from the Navy's files, and the validity coefficients they generated were somewhat higher than those produced by the Navy data. The fact that supervisors produced ratings on only one attribute or dimension, as evidenced by the limited factor structure, was no doubt a function of the fact that the scale development process was interrupted in its first phase. Had the procedures of Smith & Kendall (1963) been given a full opportunity then perhaps more interscale descrimination would have occured. At least supervisors generated enough between-worker variability that correlation with JKT scores was possible, although one correlation would have been as good as the six presented. Supervisors apparently accepted, at least in part, our assurances that the ratings would not be used for personnel actions. This was demonstrated by the rating spread between individuals.

The heart of the matter was the relationship of JKT scores and previous educational experiences. Those who had been to "A" school or had either another formal school or correspondence training scored higher on the JKT. Foley (1974) has indicated his lack of faith in paper and pencil tests, and in many ways he may have been correct. It was difficult to ascertain what percentage of the job required paper and pencil type knowledge and what percentage represented overlearned motor or perceptual habits, which had become second nature but would not lead necessarily to successful completion of a multiple choice test.

On-the-job training is primarily oriented toward performance and may not prepare an individual to answer written questions. Only a firm validation study using ultimate criteria will tease apart these variables.

However, logic dictated that paper and pencil skills would be necessary to perform many food service tests which require the use of printed materials and/or the facts listed within them. Furthermore, progression within the system required the abilities necessary to store and report on promotion tests, a wide variety of factual information. Anyone can probably learn the small subset of information necessary to do a specific task in a galley such as running a dishwashing machine or making a salad. A well rounded food specialist has to go beyond this and learn about all the jobs he may have to perform. This requires some sort of formal training.

The results of this current study were in line with the findings of previous research. Merenda (1958) found that formal school/training was superior to on-the-job training in various naval occupations when the results of promotion exams were analyzed. The close proximity in JKT scores of those who had only correspondence training to those who only had formal schooling may have been confounded by the fact that JKT items came from the correspondence programs. This leaves an open question which will require further analysis using more unique and unused test items.

The relationship of JKT scores to rank could mean that overall experiences in Navy food service do tend to enhance job knowledge of those who remain in the system. These experiences include both training and actual work performance. It was refreshing to note that many Petty Officers had superior JKT scores to those of lower pay grades, many of whom had recently completed formal training.

REFERENCES

Bureau of Naval Personnel. Commissaryman 3+2 rate training manual (NAVPERS 10279-E). Washington: U.S. Government Printing Office, 1971a.

Bureau of Naval Personnel. Steward 3+2 training manual (NAVPERS 10694-D). Washington: U.S. Government Printing Office, 1971b.

Foley, J.P. Evaluating maintenance performance: an analysis. (AFHRL-TR-74-57(1), Air Force Human Resources Laboratory, Brooks Air Force Base, Texas, Oct., 1974.

Gagne, R.M. and Fleishman, E.A. Psychology and human performance. New York: Holt, 1957.

Guilford, J.P. Psychometric methods. New York: McGraw, 1954.

Jones, L.V., Peryam, D.R. and Thurstone, L.L. Development of a scale for measuring soldiers food preferences. Food Research, 1955, 20, 512-520.

Kavanagh, M.J. The content issue in performance appraisal: a review, Personnel Psychology, 1971, 24, 653-668.

Merenda, P.F. The relative effectiveness of formal school on the job methods of training apprentices in naval occupations. Personnel Psychology, 1958, 11, 379-389.

Naval Training Command. Commissaryman 1+C rate training manual (NAVTRA 10280-F). Washington: U.S. Government Printing Office, 1973.

Naval Training Command. Steward 1+C rate training manual (NAVTRA 10695-D) Washington: U.S. Government Printing Office, 1972.

Radloff, R. and Helmreich, R. Electronic data collection in field research. American Psychologist, 1969, 24, 300-303.

Smith, P.C. and Kendall, L.M. Retranslation of expectations. An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Thorndike, R.L. Personnel Selection. New York: Wiley, 1949.

Tiffin, J. Industrial Psychology. New York: Prentice-Hall, 1952.

Vroom, V.H. Work and Motivation. New York: Wiley, 1964.

Williams, W.E. and Seiler, D.A. Supervisor and subordinate participation in the development of behaviorally anchored rating scales. Journal of Industrial and Organizational Psychology, 1973, Spring, 1-12.

Winer, B.J. Statistical principals in experimental design. New York: McGraw, 1962.

FIGURE 1

EXAMPLE OF THE PERFORMANCE SCALES
USED IN THIS STUDY

MOTIVATION

1	2	3	4	5	6	7
VERY BAD	BAD	SLIGHTLY BAD	NEITHER GOOD NOR BAD	SLIGHTLY GOOD	GOOD	VERY GOOD

CIRCLE ONE

1.	NAME 1	1	2	3	4	5	6	7
2.	NAME 2	1	2	3	4	5	6	7
3.	NAME 3	1	2	3	4	5	6	7

733

SLIDE 1

SAMPLE JOB KNOWLEDGE TEST QUESTIONS

How should the tablecloth be spread when setting the table for a meal?

a. Over the silence pad

b. Beneath the silence pad

c. Over the green felt table cover

d. On the bare table

Which of the following foods will deteriorate faster than the others?

a. Poultry

b. Seafood

c. Dairy products

d. Beef

797

SLIDE 2

CORRELATIONS BETWEEN JKT AND NAVY APTITUDE SCORES

NAVY APTITUDE TEST	CORRELATION
GENERAL CLASSIFICATION TEST	- .048
ARITHMETIC TEST	.044
MECHANICAL TEST	- .111
CLERICAL APTITUDE TEST	- .043

SLIDE 3

CORRELATIONS BETWEEN JKT AND EFFICIENCY REPORT SCORES

EFFICIENCY SCALE	CORRELATION
PERFORMANCE	.096
APPEARANCE	.056
COOPERATION	.153
RELIABILITY	.116
CONDUCT	.076
RESOURCEFULNESS	.110
LDRSHIP - DIRECTION	.134
LDRSHIP - COUNSELING	.110
OVERALL EVALUATION	.109
RECRUITER POTENTIAL	.166
INSTRUCTOR POTENTIAL	.206*
CAREER COUNSELING	.247**
INDEPENDENT DUTY	.271**

* P <.05
** P <.01

SLIDE 4

JKT Validity Using Performance Ratings

RATING	CORRELATION WITH JKT	N
Cooking Ability	.384 **	175
Supervisory Skills	.381 **	163
Motivation	.365 **	191
Personal Hygene	.315 **	191
Job Knowledge	.393 **	188
Attitude Toward Other Workers	.314 **	191

** $p < .01$

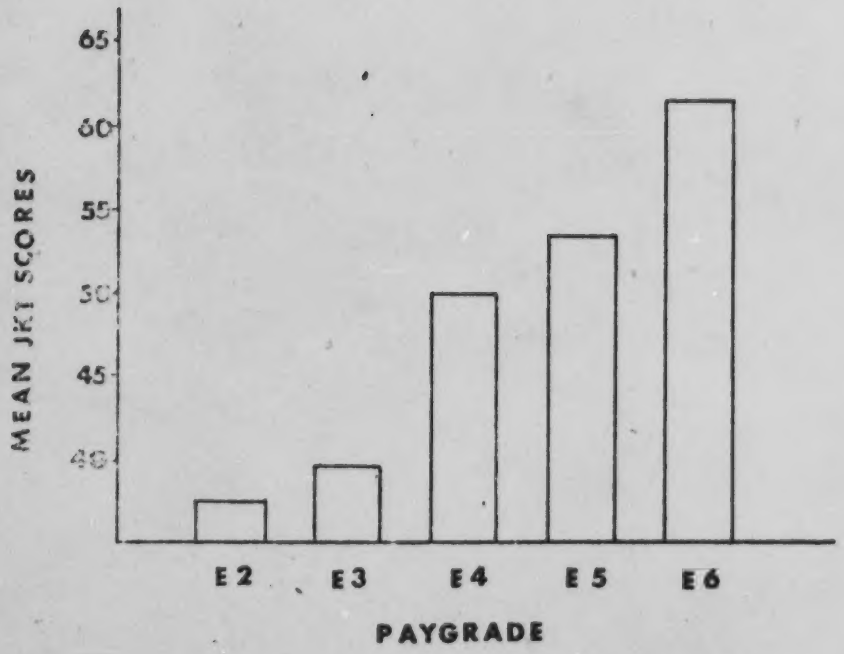
SLIDE 5

FACTOR LOADINGS ON
RATING SCALE DATA

VARIABLE	LOADINGS
COOKING ABILITY	.838
SUPERVISORY SKILLS	.823
MOTIVATION	.863
PERSONAL HYGIENE	.770
JOB KNOWLEDGE	.905
ATTITUDE TOWARD OTHER WORKERS	.824

SLIDE 6

JKT and PAYGRADE



SLIDE 7

DIFFERENCES BETWEEN AVERAGE SCORES
BASED ON PAYGRADE

GRADE		2	3	4	5	6
	MEANS	37.65	39.98	50.08	53.64	61.60
2	37.65		2.33	12.43*	15.99*	23.95*
3	39.98			10.0 *	13.66*	21.62*
4	50.08				3.56*	11.52*
5	53.64					7.96*
6	61.60					

(ANOVA F=43.1**) *P<.05, **P<.01)

GRADES SHARING COMMON UNDERLINE DO NOT DIFFER:
SCHEMATIC PRESENTATION OF DIFFERENCES.

2 3 4 5 6

SLIDE 8

RESULTS OF T TEST'S BETWEEN NO SCHOOL AND "A" GRADUATES

GRADE	T	DF	ETA²
1-3	4.33**	165	.102
4	2.00*	73	.049
5	3.54**	116	.214
6	.18	29	NA

* P <.05
** P <.01

SLIDE 9

Differences Between Average Scores for Training Categories

TRAINING	Means	NO SCHOOL	FORMAL SCHOOL ONLY	CORRESPONDANCE ONLY	BOTH FORMAL & CORRESPONDANCE
		36.9	43.23	46.07	51.42
No School	36.9		6.33*	9.17*	14.52*
Formal School Only	43.23			2.84	8.19*
Correspond. Only	46.07				5.35
Both Formal & Correspond.	51.42				

* $p < .05$

Schematic Presentation of Significant Differences

No School | Formal School | Correspondance | Formal School & Correspondance

Categories which share common underlining are not significantly different.

742

805

806

Computer-Assisted Occupational Counseling: A Prototype Model*

BY

Ted M.I. Yellen

Navy Personnel Research and Development Center
San Diego, California 92152

Traditional methods of occupationally counseling students are outdated by the state of the art in computer technology. Counseling young men and women with regard to occupational choice, be it military or civilian, is a complex and time counsuming process. It is also one of the most important decision making points in an individuals life. Unfortunetly, too often career descisions are made hastily and with little or inaccurate occupational information. To some extent this is reflected by the high attrition rate currently being experienced in the Navy. Accordingly, to achieve a better match between recruit applicants and the Navy it is becomming more and more critical to consider the total individual in terms of their career objectives, interests, attitudes, backgrounds, as well as their vocational aptitudes. The computerized counseling system described in this paper is designed to take these factors into consideration in aiding students in choosing a career path.

As part of the education program most high schools provide some form of occupational information and guidance. Some school systems have advanced to the point of providing computerized occupational guidance to their students. These counseling systems are basically Cathode Ray Tube (CRT) units in which the student provides some basic information and the system provides occupational information based upon the students input.

In many high schools the Armed Services Vocational Aptitude Battery (ASVAB) is administered to the students during their senior year. This normally is the students first actual contact with the military and in most instances their last. Although the primary purpose of the ASVAB is to stimulate the students interest in service jobs and training opportunities, most students do not enter the military service. The vast majority of students completing high school enter the labor market or continue their education.

Since most high school students desire occupational counseling and many have taken the ASVAB, the next logical step in developing a computerized counseling system was to combine civilian and military occupational information with the ASVAB as the aptitude occupational

*The opinions expressed are solely those of the author and do not necessarily reflect official Department of the Navy policy.

classification methodology. As such, the computerized occupational guidance & counseling system would provide students with civilian as well as Navy occupations. These jobs would be selected on the basis of measured aptitude as well as the individuals personal interests and job characteris[illegible] preferences.

The procedure [illegible] tablishing equivalant aptitude measurement scores between civil[illegible] jobs and corresponding ASVAB scores is contained in a report entitled "A computerized counseling system for civilian and related Navy occupations" (Yellen, in preparation).

One of the system's characteristics is the degree to which existing information retrieval systems have been modified into a unified civilian and Navy counseling system. As shown in Figure 1 the system incorporated job classification features from three civilian occupational information systems, and the Navys' school eligibility requirements.

The system contains 279 civilian jobs, 114 worker trait groups, 76 basic Navy occupations, and over 100 specialized Navy jobs. It is a real-time interactive system and a prototype of a future, expanded Navy occupational counseling system. The present system is designed primarily for use by high school students. The system matches their attributes with occupational requirements. Although the system provides occupational information concerning Navy ratings its main objective is to acquaint high school students with various civilian careers which they may want to explore. It is anticipated that the majority of students are interested only in the civilian job information. However, all students using the system will see, probably for the first time, the relationship between civilian jobs and Navy occupations and also which training programs they are qualified to enter. For those who desire more information concerning the Navy jobs, the expanded version of the counseling program will be available for their use. The subsequent counseling dialogues will deal with Navy job opportunities, training schools, promotion rates, duty assignments and the like.

The occupational information exchange in the prototype system is managed by the computer and the student. With the possible exception of turning on the CRT unit no other assistance is required to operate the system. All that is required of the students is to enter their ASVAB scores and some occupational like and dislike information into the system. The requested information and instructions for operating the system are contained in a guidance and counseling booklet. Also, the computer program leads the student through all the information exchange dialogue.

The counseling booklet contains 31 questions, 56 school subjects, and 279 specific jobs. As shown in Figure 1 the personal input information consists of occupational interest areas, educational goals, aptitudes (ASVAB), job characteristic preferences, and school subject preferences. The students record their responses to the questions on an answer sheet and then enter this information into the computer via

a CRT terminal. A dialogue between the computer and the student follows. This involves discussing some of the information the student entered and allows the student to make changes if they desire. After the changes are made the computer provides the student with a printout containing the information shown in Table 1.

Table 1

Occupational Information Output

- Identifies civilian jobs that match the aptitudes, occupational interests, and educational plans of the individual involved.
- Provides a short description for each selected civilian job.
- Provides a descriptive comparison between the individuals personal traits with those associated with the selected jobs.
- Identifies inconsistencies among educational plans, occupational interests, aptitudes, stated career choices, and occupational requirements.
- Identifies related Navy occupations for each of the selected civilian jobs.
- Provides Navy percentile scores for each aptitude subtest and a summary of the questionnaire responses.
- Identifies the Navy occupations for which the individual is qualified for entry on the basis of ASVAB scores.
- Identifies worker trait groups[1] that correspond to the individuals school subject interests, aptitudes (as measured by the ASVAB), interest factors (type of work activities) and temperament (ability to adjust or cope with various work situations).
- Describes the work performed for each worker trait groups and explains the training and methods of entry into the field.
- Provides examples of the civilian jobs contained in each worker trait group and identifies the related Navy occupations for the civilian jobs.
- Provides sources for obtaining additional occupational information.

[1] A classification system contained in the Dictionary of Occupational Titles in which civilian jobs possessing similar worker characteristics are catergorized into 114 worker trait groups.

An example of a computerized counseling printout is provided in the following pages. As shown on page 7, this summary report is for a high school senior named Harry Student.

Presented on page 8 are the raw ASVAB scores that Harry entered into the computer. Also recorded are his percentile scores. As shown on pages 8 and 9, the printout recaps Harrys' input information regarding his educational plans, interest areas, job characteristics preferences, preferred working situations, school subject preferences and his specific occupational choices.

In section I of the printout (see pages 9 thru 15) Harry is provided with a list of civilian jobs that match his aptitudes, occupational interests, and educational plans. The 279 jobs are coded according to these criterion factors. On the basis of his aptitude scores (as measured by the ASVAB), occupational interests (as determined by the Kuder Occupational Interest Inventory categories), and his stated educational plans, the computer screens the 279 civilian jobs and selects those which match the coded job requirements. For example, the first job on page 10 - Photographic Laboratory workers - enter into this job typically does not require a high school diploma. The job is catcgorized as "Clerical" in the occupational interest area. With the exception of space relations, the aptitude levels normally required to understand and learn the principles of operation for this job are below average for Mechanical, Clerical, Arthmetic Reasoning and Word Knowledge. Ability to perceive real objects and relationships from pictures is considered important for success and accordingly the space relations scores should be average or better.

Below the job title a short description of each job is provided. Notice that each job is described in terms of a number of job characteristics and requirements. In the counseling booklet answer sheet the students indicated their preferences for working under different conditions. Each of the 279 jobs is described by some combination of these characteristics and requirements. For example, if a characteristic of a job requires working independently, a question in the booklet attempts to determine and individuals preference for working independetly or as part of a team. The computer matches the occupational characteristics and requirements with the student's personal characteristic preferences. As you can see the printout lists the characteristics and requirements applicable to the selected jobs and describes each as being positive, negative, or as neutral features for the individual. In this example, Harry prefers working a part of a team. Accordingly, for Harry, working independently is a negative feature of this job. The printout also identifies the related Navy occupations for each of the selected civilian jobs. If the student is not ASVAB qualified for a related Navy job, it will not be presented.

As shown on the bottom of page 15 the printout identified some inconsistencies regarding the Harrys input information and the job requirements. In this example, Harry selected accountant as a job he

may like to enter. The printout indicated that this job was not selected as a possible career choice because Harrys' ASVAB scores were too low and also his educational plans were inconsistant with the job requirements. Harry stated that he is planning on completing high school. Typically, to get a job as accountant requires the completion of a 4 year college program. For these reasons this job was not selected as one for Harry's consideration.

In Section II of the printout (see page 15) Harry is presented with a list of all the Navy occupations for which he is qualified to enter on the basis of his ASVAB scores. Of course entry into these job's are subject to the needs of the Navy at any particular time.

In Section III (see pages 16 thru 21) Harry is presented with a number of worker trait groups that correspond to his school subject interests, aptitudes, interest factors and temperament (ability to adjust or cope with various work situations).

The printout describes the work performed for each of the worker trait groups and explains the training and methods of entry into the field. Below this explanation are examples of the civilian jobs contained in each worker trait group and the related Navy occupations. It is noted in the explanations that although the related Navy jobs are presented in this section the individual may not be qualified for entry into them.

As shown on page 17 Harry is provided with a list of sources for obtaining additional occupational information. Also, in the final portion of the printout (see page 21) guidlines are provided outlining the steps he may want to take concerning his career decision.

This is the first stage of several counseling dialogues being developed. This one is, by design, limited in the amount of Navy information provided to the student. As previously mentioned, for those students interested in obtaining more information about the Navy an expanded version of the counseling dialogue will be available. As for implementation, we plan to field test the counseling system in several high schools located in the San Diego and Los Angeles school districts. The Center has a mobile van which is designed to accomodate 7 CRT units. Accordingly the self containted van can be parked at a high school and provide occupational counseling to the students with minimal distrubance.

A detailed explanation of the procedure for implementing the counseling system will be discussed in the following paper entitled "The selection of an occupational counseling computer netowrk system for the Navy recruiting environment.

```
 CCCCC    OOOOO   MM     MM  PPPPPP   U     U  TTTTTTT  EEEEEEE  RRRRRR   III  ZZZZZZZ  EEEEEEE  DDDDD
C     C  O     O  M M   M M  P     P  U     U     T     E        R     R   I        Z   E        D    D
C        O     O  M  M M  M  P     P  U     U     T     E        R     R   I       Z    E        D     D
C        O     O  M   M   M  PPPPPP   U     U     T     EEEEE    RRRRRR    I      Z     EEEEE    D     D
C        O     O  M       M  P        U     U     T     E        R   R     I     Z      E        D     D
C     C  O     O  M       M  P        U     U     T     E        R    R    I    Z       E        D    D
 CCCCC    OOOOO   M       M  P         UUU      T     EEEEEEE  R     R  III  ZZZZZZZ  EEEEEEE  DDDDD

 OOOOO    CCCCC    CCCCC   U     U  PPPPPP     AAA     TTTTTTT  III   OOOOO   N     N    AAA    L
O     O  C     C  C     C  U     U  P     P   A   A       T      I   O     O  NN    N   A   A   L
O     O  C        C        U     U  P     P  A     A      T      I   O     O  N N   N  A     A  L
O     O  C        C        U     U  PPPPPP   A     A      T      I   O     O  N  N  N  A     A  L
O     O  C        C        U     U  P        AAAAAAA      T      I   O     O  N   N N  AAAAAAA  L
O     O  C     C  C     C  U     U  P        A     A      T      I   O     O  N    NN  A     A  L
 OOOOO    CCCCC    CCCCC    UUU    P        A     A      T     III   OOOOO   N     N  A     A  LLLLLLL

 CCCCC    OOOOO   U     U  N     N   SSSSS   EEEEEEE  L         III  N     N   GGGGG
C     C  O     O  U     U  NN    N  S     S  E        L          I   NN    N  G     G
C        O     O  U     U  N N   N  S        E        L          I   N N   N  G     G
C        O     O  U     U  N  N  N   SSSSS   EEEEE    L          I   N  N  N  G
C        O     O  U     U  N   N N        S  E        L          I   N   N N  G   GGG
C     C  O     O  U     U  N    NN  S     S  E        L          I   N    NN  G     G
 CCCCC    OOOOO    UUU    N     N   SSSSS   EEEEEEE  LLLLLLL  III  N     N   GGGGG
```

748

SUMMARY REPORT
FOR

****
* HARRY STUDENT *
****

DEVELOPED BY THE NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIF
(IN SUPPORT OF THE NAVY RECRUITING COMMAND WASHINGTON, D.C.)

812

813

SUMMARY INFORMATION (ASVAB)

SCORE	GENERAL INFORMATION	NUMERICAL OPERATION	ATTENTION TO DETAIL	WORD KNOWLEDGE	ARITHMETIC REASONING	SPACE PERCEPTION	MATHEMATIC KNOWLEDGE	ELECTRONIC INFORMATION
ASVAB#	10	34	14	27	8	12	9	16
PERCENTILE@	53	60	50	90	24	62	45	44

SCORE	MECHANICAL COMPREHENSION	GENERAL SCIENCE	SHOP INFORMATION	AUTOMOTIVE INFORMATION
ASVAB#	8	10	12	10
PERCENTILE@	37	47	45	45

ASVAB RAW SCORE
@ PERCENTILE SCORES ARE BASED UPON A MILITARY APPLICANT SAMPLE. THOSE REPORTED IN THE COUNSELOR PRINTOUT GUIDE ARE SCHOOL AND NATION-WIDE PERCENTILES.

749

SUMMARY INFORMATION (EDUCATION & INTEREST AREAS)

EDUCATION . . . . . YOU INDICATED THAT YOU INTEND TO GRADUATE FROM HIGH SCHOOL.

INTERESTS. . . . . YOU INDICATED THE TWO TYPES OF WORK THAT ARE OF MOST INTEREST TO YOU AT THIS TIME ARE:

COMPUTATIONAL
CLERICAL

SUMMARY INFORMATION (JOB CHARACTERISTICS)

JOB CHARACTERISTICS: THE FOLLOWING JOB CHARACTERISTICS ARE THOSE YOU SELECTED AS THOSE YOU WOULD LIKE IN A CAREER:

PREFERED CHARACTERISTICS	DISLIKED CHARACTERISTICS	CHARACTERISTICS I NOT SURE ABOUT
5. RESPONSIBILITY IN MAKING DECISIONS AND EXERCISING GOOD JUDGEMENT.	11. REQUIRES PHYSICAL STAMINA.	9. HELPING PEOPLE.
	13. WORKING WITH IDEAS.	12. OVERTIME OR SHIFT WORK.
SUPERVISORY RESPONSIBILITY IN	17. EXPOSURE TO WEATHER.	14. WORKING WITH PEOPLE.

814

815

DIRECTING ACTIVITIES OF OTHERS.

7. ROUTINE AND REPETITIOUS JOBS.

8B. FOLLOW A SET PROCEDURE.

10A. WORKING AS PART OF A TEAM.

15. WORKING WITH THINGS.

16. CONFINED TO WORK AREA.

18. WORKING WITH DETAILS.

19. WORKING UNDER CLOSE SUPERVISION.

21. REQUIRES YOU TO MOTIVATE OTHERS.

22. WORKING UNDER STRESS.

20. HIGHLY COMPETITIVE JOB.

23. ARTISTIC WORK.

PREFERED WORKING SITUATIONS:

24. CONCERNED WITH PEOPLE AND THE COMMUNICATION OF IDEAS.
25. BUSINESS CONTACT WITH PEOPLE
26. A ROUTINE, CONCRETE, ORGANIZED NATURE.
27. ACTIVITIES THAT ARE NONSOCIAL.
28. HANDS ON TYPE WORK.
29. REPETITIVE OF SHORT CYCLE OPERATIONS CARRIED OUT ACCORDING TO SET PROCEDURES OR SEQUENCES.
30. DOING THINGS UNDER SPECFIC INSTRUCTION.
31. EXCHANGING IDEAS, AND INFORMATION WHILE DEALING WITH PEOPLE IN ACTUAL JOB DUTIES.

SCHOOL SUBJECT PREFERENCES. . . . . YOU INDICATED THAT YOU LIKE THE FOLLOWING SCHOOL SUBJECTS:

750

11. EARTH SCIENCES
13. GOVERNMENT, CIVICS
14. ECONOMICS
18. PHYSICAL EDUCATION
19. SPORTS, RECREATION
24. INDUSTRIAL ARTS
32. FINANCE, REAL ESTATE, INSURANCE
33. DENTAL-MEDICAL ASSISTANT
40. ACCOUNTING, BOOKKEEPING
48. COMMERCIAL ART

OCCUPATIONAL CHOICE. . . .YOU INDICATED THAT THE CIVILIAN JOB YOU WOULD LIKE TO ENTER ARE:

JOB NUMBER 34 WHICH IS: BOOKKEEPING WORKERS
JOB NUMBER 56 WHICH IS: ACCOUNTANTS

817

816

SECTION I -CIVILIAN JOBS-

IE COMPUTER HAS IDENTIFIED THE FOLLOWING CIVILIAN JOBS AS THOSE MOST SUITED FOR YOU ON THE BASIS OF THE INFORMATION
OU PROVIDED ON YOUR ASVAB SCORES, THE HIGHEST LEVEL OF EDUCATION YOU PLAN TO COMPLETE AND YOUR STATED OCCUPATIONAL

INTERESTS. IT SHOULD BE NOTED THAT THE SET OF JOBS COULD BE DIFFERENT IF YOU ENTERED OTHER THAN THE INFORMATION PROVIDED. THE DEGREE OF AGREEMENT BETWEEN YOUR LIKES AND DISLIKES AND THE PARTICULAR JOB CHARACTERISTICS AND REQUIREMENTS ARE SHOWN FOR EACH SELECTED JOB. THIS INFORMATION IS PROVIDED TO HELP YOU DETERMINE IF YOU SHOULD PURSUE THE JOB AS A POSSIBLE CAREER CHOICE. IF THERE ARE ONE OF TWO CHARACTERISTICE ABOUT A JOB THAT -TURN YOU OFF- YOU PROBABLY SHOULD CONSIDER ANOTHER OCCUPATION. ALSO SHOWN ARE SOME OF THE RELATED NAVY OCCUPATIONS. YOU MAY WANT TO INVESTIGATE THESE AS POSSIBILITIES FOR IMMEDIATE EMPLOYMENT. THE TECHNICAL TRAINING RECEIVED AND WORK PERFORMED HAVE DIRECT APPLICATION TO THE CIVILIAN JOBS.

****
****

PHOTOGRAPHIC LABORATORY WORKER:
PHOTOGRAPHIC LABORATORY OCCUPATIONS INCLUDE WORKERS WHO DEVELOP FILM, MAKE PRINTS AND SLIDES, AND PERFORM RELATED TASKS SUCH AS ENLARGING PHOTOGRAPHS.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:	FOR YOU THIS IS A: **********
REPETITIOUS WORK	POSITIVE FEATURE
WORKING INDEPENDENTLY	NEGATIVE FEATURE
WORKING WITH THINGS	POSITIVE FEATURE
GENERALLY CONFINED TO A WORK AREA	POSITIVE FEATURE
WORKING WITH DETAILS	POSITIVE FEATURE
JOB SATISFACTION OBTAINED FROM SEEING PHYSICAL RESULTS OF WORK	POSTIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

751

LITHOGRAPHER
PHOTOGRAPHER'S MATE

****
****
****

BOOKKEEPING WORKERS:
BOOKKEEPING WORKERS MAINTAIN RECORDS OF DAY-TO-DAY BUSINESS TRANSACTIONS IN JOURNALS AND LEDGERS AND ON OTHER ACCOUNTING FORMS.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:	FOR YOU THIS IS A: **********
REPETITIOUS WORK	POSITIVE FEATURE
WORKING AS PART OF A TEAM	POSITIVE FEATURE
GENERALLY CONFINED TO A WORK AREA	POSITIVE FEATURE
WORKING WITH DETAILS	POSITIVE FEATURE
WORKING UNDER CLOSE SUPERVISION	NEGATIVE FEATURE

819

818

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

AVIATION MAINTENANCE ADMINISTRATIONMAN

AVIATION STOREKEEPER
DISBURSING CLERK
STOREKEEPER
SHIP'S SERVICEMAN

****
****
****

CASHIERS:
CASHIERS ACCEPT MONEY PAID BY CUSTOMERS AND CLIENTS, MAKE CHANGE WHEN NECESSARY, AND GIVE SOME KIND OF RECEIPT FOR PAYMENT AND THEY KEEP RECORDS OF THE AMOUNT OF MONEY INVOLVED IN EACH TRANSACTION SO THAT CASH ACCOUNTS CAN BE BALANCED AT THE END OF THE DAY.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES: — FOR YOU THIS IS A:
****

REPETITIOUS WORK.....POSITIVE FEATURE
WORKING INDEPENDENTLY.....NEGATIVE FEATURE
WORKING WITH PEOPLE.....NEUTRAL FEATURE
GENERALLY CONFINED TO A WORK AREA.....POSITIVE FEATURE
WORKING WITH DETAILS.....POSITIVE FEATURE
WORKING UNDER CLOSE SUPERVISION.....NEGATIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

AVIATION STOREKEEPER
DISBURSING CLERK
SHIP'S SERVICEMAN
STOREKEEPER

752

****
****
****

FILE CLERKS:
FILE CLERKS PRODUCE TYPED COPIES OF PRINTED AND HANDWRITTEN MATERIALS OR TYPE DICTATION RECORDED ON TAPE OR ON SOUND-PRODUCING RECORDS.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES: — FOR YOU THIS IS A:
****

REPETITIOUS WORK.....POSITIVE FEATURE
WORKING AS PART OF A TEAM.....POSITIVE FEATURE
WORKING WITH IDEAS.....NEGATIVE FEATURE
WORKING WITH DETAILS.....POSITIVE FEATURE
WORKING UNDER CLOSE SUPERVISION.....NEGATIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

AVIATION MAINTENANCE ADMINISTRATIONMAN
AVIATION STOREKEEPER
DISBURSING CLERK
HOSPITAL CORPSMAN

820

SHIP'S SERVICEMAN
STOREKEEPER
YEOMAN

****
****
****

OFFICE MACHINE OPERATORS:
OFFICE MACHINE OPERATORS WORK ON MACHINES THAT RANGE FROM SIMPLE MECHANICAL DEVICES THAT OPEN LETTERS TO ELECTRONIC EQUIPMENT CAPABLE OF PERFORMING HIGHLY INVOLVED COMPUTATIONS.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:	FOR YOU THIS IS A: **********
REPETITIOUS WORK	POSITIVE FEATURE
WORKING AS PART OF A TEAM	POSITIVE FEATURE
WORKING WITH THINGS	POSITIVE FEATURE
GENERALLY CONFINED TO A WORK AREA	POSITIVE FEATURE
WORKING WITH DETAILS	POSITIVE FEATURE
WORKING UNDER CLOSE SUPERVISION	NEGATIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

STOREKEEPER

****
****
****

753

STOCK CLERKS:
ONE WHO IS CONCERNED WITH KEEPING RECORDS AND TRACK OF THE STATE OF SUPPLIES AND/OR MERCHANDISE.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:	FOR YOU THIS IS A: **********
WORKING AS PART OF A TEAM	POSITIVE FEATURE
PHYSICAL STAMINA	NEGATIVE FEATURE
WORKING WITH THINGS	POSITIVE FEATURE
WORKING UNDER CLOSE SUPERVISION	NEGATIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

AVIATION MAINTENANCE ADMINISTRATIONMAN
AVIATION STOREKEEPER
SHIP'S SERVICEMAN
STOREKEEPER

****
****
****

821

GENERAL STENOGRAPHERS TAKE FAIRLY ROUTINE DICTATION AND PERFORM ROUTINE OFFICE TASKS.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:	FOR YOU THIS IS A:
REPETITIOUS WORK	POSITIVE FEATURE
WORKING INDEPENDENTLY	NEGATIVE FEATURE
WORKING WITH THINGS	POSITIVE FEATURE
GENERALLY CONFINED TO A WORK AREA	POSITIVE FEATURE
WORKING WITH DETAILS	POSITIVE FEATURE
WORKING UNDER CLOSE SUPERVISION	NEGATIVE FEATURE
JOB SATISFACTION OBTAINED FROM SEEING PHYSICAL RESULTS OF WORK	POSTIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

LEGELMAN, FROM THE YEOMAN RATING
YEOMAN

****
****
****

TYPISTS:
TYPISTS PRODUCE TYPED COPIES OF PRINTED AND HANDWRITTEN MATERIALS OR TYPE DICTATION RECORDED ON TAPE OR ON SOUND-PRODUCING RECORDS.

754

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:	FOR YOU THIS IS A:
REPETITIOUS WORK	POSITIVE FEATURE
WORKING INDEPENDENTLY	NEGATIVE FEATURE
WORKING WITH THINGS	POSITIVE FEATURE
GENERALLY CONFINED TO A WORK AREA	POSITIVE FEATURE
WORKING WITH DETAILS	POSITIVE FEATURE
WORKING UNDER CLOSE SUPERVISION	NEGATIVE FEATURE
JOB SATISFACTION OBTAINED FROM SEEING PHYSICAL RESULTS OF WORK	POSTIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

LEGELMAN, FROM THE YEOMAN RATING
YEOMAN

****
****
****

CLAIM EXAMINERS:
PERSONS CHARGED WITH THE INSPECTION OF PERSONS ENTERING AND LEAVING
NATION TO INSURE COMPLIANCE WITH THE LAWS OF THAT NATION.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:

FOR YOU THIS IS A:
****

DIRECTING THE ACTIVITIES OF OTHERS.....POSITIVE FEATURE
REPETITIOUS WORK.....POSITIVE FEATURE
WORKING INDEPENDENTLY.....NEGATIVE FEATURE
WORKING WITH IDEAS.....NEGATIVE FEATURE
BEING EXPOSED TO WEATHER CONDITIONS.....NEGATIVE FEATURE
WORKING WITH DETAILS.....POSITIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

DISBURSING CLERK
LEGELMAN, FROM THE YEOMAN RATING
MASTER-AT-ARMS, MUST BE A 1ST CLASS

****
****
****

TRAFFIC AGENTS AND CLERKS:
ONE WHO SELLS TICKETS FOR AIR FLIGHTS, AND WHO MAKES RESERVATIONS.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:

FOR YOU THIS IS A:
****

755

HELPING PEOPLE.....NEUTRAL FEATURE
WORKING INDEPENDENTLY.....NEGATIVE FEATURE
OVERTIME OR SHIFT WORK.....NEUTRAL FEATURE
WORKING WITH PEOPLE.....NEUTRAL FEATURE
GENERALLY CONFINED TO A WORK AREA.....POSITIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

AVIATION STOREKEEPER
DISBURSING CLERK
STOREKEEPER
YEOMAN

****
****
****

CONDUCTORS:
CONDUCTORS ARE RESPONSIBLE FOR THE SAFETY OF THEIR PASSENGERS AND CARGOES, AND THEY SUPERVISE THE WORK OF THE TRAIN AND ENGINE CREWS.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:

FOR YOU THIS IS A:
****

823

DIRECTING THE ACTIVITIES OF OTHERS.....POSITIVE FEATURE
REPETITIOUS WORK.....POSITIVE FEATURE
HELPING PEOPLE.....NEUTRAL FEATURE
PHYSICAL STAMINA.....NEGATIVE FEATURE

WORKING WITH PEOPLE.....NEUTRAL FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

THERE ARE NO RELATED NAVY ENLISTED CAREERS.

****
****
****

TELEGRAPHERS AND TOWER WORKERS:
ONE WHO WORKS IN THE TOWER ON THE RAILROAD FOR DIRECTING TRAFFIC.

THIS OCCUPATION TYPICALLY PROVIDES OR REQUIRES:	FOR YOU THIS IS A:
	**********
WORKING AS PART OF A TEAM.....	POSITIVE FEATURE
OVERTIME OR SHIFT WORK.....	NEUTRAL FEATURE
WORKING WITH THINGS.....	POSITIVE FEATURE
GENERALLY CONFINED TO A WORK AREA.....	POSITIVE FEATURE
WORKING UNDER CLOSE SUPERVISION.....	NEGATIVE FEATURE

YOUR ASVAB SCORES QUALIFY YOU FOR THE FOLLOWING RELATED NAVY CAREERS:

COMMUNICATIONS TECHNICIAN
RADIOMAN

****

756

THE JOB - ACCOUNTANTS
WAS NOT SELECTED AS A POSSIBLE CAREER CHOICE BECAUSE:

YOUR ASVAB SCORES ARE TOO LOW FOR THIS JOB.

YOUR EDUCATIONAL PLANS ARE INCONSISTENT WITH THE JOB REQUIREMENTS. THIS JOB NORMALLY REQUIRES:
A 4 YEAR COLLEGE DEGREE.

SECTION II -NAVY OCCUPATIONS-

YOU QUALIFY FOR THE FOLLOWING NAVY OCCUPATIONS:

PR AIRCREW SURVIVAL EQUIPMENTMAN
AB AVIATION BOATSWAIN'S MATE
AD AVIATION MACHINIST'S MATE
AZ AVIATION MAINTENANCE ADMINISTRATIONMAN
AK AVIATION STOREKEEPER
AM AVIATION STRUCTURAL MECHANIC
AS AVIATION SUPPORT EQUIPMENT TECHNICIAN
BM BOATSWAIN'S MATE
BO BOILER MAKER. MUST BE A 1ST CLASS BOATSWAIN'S MATE

824

BT BOILER TECHNICIAN
BU BUILDER
CTA COMMUNICATIONS TECHNICIAN, ADMINISTRATIVE
CTI COMMUNICATIONS TECHNICIAN, INTERPRETIVE, NEED 23 ON FLAT TEST
CTO COMMUNICATIONS TECHNICIAN, COMMUNICATIONS
CTR COMMUNICATIONS TECHNICIAN, COLLECTION, NEED A 60 ON RADIO TEST
CTT COMMUNICATIONS TECHNICIAN, TECHNICAL
CE CONSTRUCTION ELECTRICIAN
CM CONSTRUCTION MECHANIC
DT DENTAL TECHNICIAN
DK DISBURSING CLERK
EM ELECTRICIAN'S MATE
EA ENGINEERING AID
EN ENGINEMAN
EO EQUIPMENT OPERATOR
GS GAS TURBINE SYSTEM TECHNICIAN, MUST BE A 1ST CLASS ENGINEMEN
HM HOSPITAL CORPSMAN
HT HULL MAINTENANCE TECHNICIAN
DM ILLUSTRATOR-DRAFTSMAN
IC INTERIOR COMMUNICATIONS ELECTRICIAN
JO JOURNALIST
LN LEGELMAN, FROM THE YEOMAN RATING
MR MACHINERY REPAIRMAN
LI LITHOGRAPHER
MM MACHINIST'S MATE
MA MASTER-AT-ARMS, MUST BE A 1ST CLASS
MN MINEMAN
ML MOLDER
MU MUSICIAN (AUDITIONS REQUIRED)
NC NAVY COUNSELOR, MUST BE A 1ST CLASS
PM PATTERNMAKER
IS INTELLIGENCE SPECIALIST
PH PHOTOGRAPHER'S MATE
QM QUARTERMASTER
RM RADIOMAN
SH SHIP'S SERVICEMAN
SM SIGNALMAN
SW STEELWORKER
MS MESS SPECIALIST
SK STOREKEEPER
UT UTILITIESMAN
YN YEOMAN

757

SECTION III

EXPLANATION OF OCCUPATIONAL INFORMATION PRESENTED IN THIS SECTION

THE CIVILIAN JOBS PRESENTED IN THE FOLLOWING PAGES WERE SELECTED FOR YOU ON THE BASIS OF A SEPERATE SCREENING. THAT IS, A DIFFERENT SET OF 'INPUT INFORMATION' (QUESTIONS IN THE BOOKLET) WAS USED TO INDENTIFY THESE JOBS. ACCORDINGLY THEY MAY NOT CORRESPOND TO THOSE PREVIOUSLY SELECTED FOR YOU.

IN THIS OCCUPATIONAL SEARCH CIVILIAN JOBS THAT HAVE THE SAME BASIC OCCUPATIONAL, INDUSTRIAL, AND WORKER CHARACTERISTIC REQUIREMENTS ARE GROUPED TOGETHER. THESE ARE CATEOGORIZED INTO 114 DIFFERENT WORKER TRAIT GROUPS. YOUR SCHOOL

AND [illegible] SCORES WERE COMPARED WITH OF THE WORKER TRAIT GROUP REQUIREMENTS. I FOUND 7 WORKER TRAIT GROUPS WHICH WERE IN COMPLETE AGREEMENT. THE FOLLOWING PAGES CONTAIN GENERAL DESCRIPTIONS OF THE WORK PERFORMED IN EACH THESE WORKER TRAIT GROUPS, THE TRAINING AND METHODS OF ENTRY INTO THE JOBS, AND EXAMPLES OF CIVILIAN JOBS CONTAINED IN THESE GROUPS. IN ADDITION, WHEREVER POSSIBLE THE RELATED NAVY OCCUPATIONS ARE LISTED ALONG SIDE THEIR CIVILIAN COUNTERPARTS. SOME OF THE NAVY OCCUPATIONS HAVE A NUMBERED CODE IN PARENTHESIS AFTER THE OCCUPATIONAL TITLE. THE CODE INDICATES THAT THESE OCCUPATIONS REQUIRE SPECIALIZED TRAINING AND OR EXPERIENCE. ACCORDINGLY, ENTRY INTO THESE NAVY OCCUPATIONS REQUIRES MEETING THE BASIC OCCUPATIONAL APTITUDE REQUIREMENTS AND, IN MOST CASES, ADDITIONAL NAVY SCHOOL TRAINING. THE LETTERS IN THE PARENTHESIS INDICATE THE BASIC OCCUPATIONAL ENTRY LEVEL. ON THE BASIS OF YOUR ASVAB SCORES A LIST OF THE BASIC NAVY OCCUPATIONS FOR WHICH YOU ARE QUALIFIED TO ENTER IS PROVIDED IN SECTION II OF THIS PRINTOUT.

SOURCES OF ADDITIONAL INFORMATION

FOR MORE INFORMATION ABOUT THESE AND OTHER JOBS, SEE:

(1) VOLUMES I AND II OF THE 'DICTIONARY OF OCCUPATIONAL TITLES' (DOT), THIRD EDITION. VOLUME I OF THE DOT FURNISHES JOB DESCRIPTIONS AND VOLUME II PROVIDES ADDITIONAL JOBS FOR EACH OF THE WORKER TRAIT GROUPS.

(2) 'OCCUPATIONAL OUTLOOK HANDBOOK' PROVIDES JOB DESCRIPTIONS AND ALSO ANSWERS SUCH QUESTIONS AS: WHAT DOES THE PERSON ACTUALLY DO ON THE JOB; WHAT ABILITIES AND INTERESTS DOES THE JOB CALL FOR; WHAT KIND OF SCHOOLING AND OTHER TRAINING IS REQUIRED FOR THE JOB; WHAT ARE THE WORKING CONDITIONS LIKE; AND WHAT WILL BE THE JOB OPPORTUNITIES IN THE COMING YEARS.

(3) FOR SPECIFIC INFORMATION ABOUT THE NAVY OCCUPATIONS, THE LOCAL NAVY RECRUITING OFFICE SHOULD BE CONTACTED.

758

WORKER TRAIT GROUP: INTERVIEWING, INFORMATION-GIVING, AND RELATED WORK (VOCATIONAL, EDUCATIONAL, AND RELATED ACTIVITIES) (DOT VOL. II PAGE 250)

WORK PERFORMED: WORK ACTIVITIES IN THIS GROUP PRIMARILY INVOLVE INTERVIEWING INDIVIDUALS IN ORDER TO GATHER PERTINENT INFORMATION AND EVALUATE THE INFORMATION AND/OR THE INDIVIDUAL'S QUALIFICATIONS FOR A VARIETY OF CONSIDERATIONS, AND DISPENSING INFORMATION USUALLY RELATIVE TO THE CORRECT INTERPRETATION OF RULES AND REGULATIONS GOVERNING SUCH MATTERS AS INSURANCE, EDUCATION, OR HOUSING.

TRAINING AND METHODS OF ENTRY: A COLLEGE EDUCATION IS BECOMING INCREASING IMPORTANT FOR ENTRANCE INTO WORK OF THIS NATURE. COURSES IN PERSONNEL MANAGEMENT, BUSINESS ADMINISTRATION, LIBERAL ARTS, APPLIED PSYCHOLOGY, AND SOCIOLOGY FREQUENTLY WILL ENABLE THE APPLICANT TO OBTAIN SUCCESSFUL EMPLOYMENT. OFTEN HOWEVER COMPANIES WILL FILL POSTIONS BY OFFERING ON-THE-JOB TRAINING TO PERSONAL ALREADY EMPLOYED WHO ARE FAMILIAR WITH COMPANY PROCEDURES.

EXAMPLES OF CIVILIAN JOBS: RELATED NAVY OCCUPATIONS:

EMPLOYMENT INTERVIEWER

PRISONER-CLASSIFICATION INTERVIEWER

PERSONNELMAN (PN)
CLASSIFICATION INTERVIEWER (PN-2612)

CORRECTIONAL COUNSELER (PN-2616)

****

WORKER TRAIT GROUP: TITLE AND CONTRACT SEARCH AND ANAYLSIS AND RELATED WORK
(DOT VOL. II PAGE 254)

WORK PERFORMED: WORK ACTIVITIES IN THIS GROUP PRIMARILY INVOLVE EXAMINING, AUTHENTICATION, AND PREPARING SUCH LEGAL AND BUSINESS DOCUMENTS AS TITLES, CONTRACTS, MORTGAGES, AND LIENS.

TRAINING AMD METHODS OF ENTRY: ENTRY INTO THIS KIND OF WORK FREQUENTLY REQUIRES THE COMPLETION OF COURSEWORK IN LAW OR BUSINESS AMD FAMILIARITY WITH THE SPECIFIC TYPE OF ACTIVITY INVOLVED. A COLLEGE DEGREE OR CONSIDERABLE COLLEGE OR NIGHT SCHOOL COURSE WORK WITH AN EMPHASES ON BUSINESS OR LAW PROVIDES EXCELLENT PREPARATION FOR ENTRY. EMPLOYERS WILL OFTEN CONSIDER THE INDIVIDUAL WHO HAS ACHIEVED A HIGH DEGREE OF PROFICIENCY AT LESS COMPLEX WORK IN THE SAME ACTIVITY, PROVIDED HE HAS BEEN EXPOSED TO DOCUMENTARY TERMINOLOGY AND REGULATIONS.

THE BASIC REQUISITES FOR A CIVILIAN CAREER IN THIS FIELD MAY BE ACQUIRED BY STUDING FOR THE RATINGS OF 'LEGALMAN' IN THE NAVY.

EXAMPLES OF CIVILIAN JOBS:	RELATED NAVY OCCUPATION:
**********	**********
CLAIM EXAMINER	LEGALMAN (LN)
TITLE CLERK	
LAW CLERK	

****

WORKER TRAIT GROUP: CORRESPONDING AND RELATED WORK
(DOT VOL. II PAGE 256)

WORK PERFORMED: WORK ACTIVITIES IN THIS GROUP PRIMARILY INVOLVE PREPARING CORRESPONDENCE, REPORTS, FORMS, AND OTHER DOCUMENTS, AND EXAMINING, EDITING, AND RECORDING VERBAL INFORMATION OR WRITTEN MATERIAL.

TRAINING AND METHODS OF ENTRY: ENTRY INTO THIS KIND OF WORK USUALLY REQUIRES HIGH SCHOOL GRADUATION AND COMPLETION OF SEVERAL SEMESTERS OF COLLEGE WITH EMPHASIS ON COURSES IN ENGLISH GRAMMER AND COMPOSITION.

ENTRY IS FREQUENTLY ACCOMPLISHED BY ADVANCEMENT OR TRANFER FROM CLERICAL POSITIONS WHEREIN INDIVIDUALS HAVE BECOME FAMILIAR WITH CORRESPONENCE TECHNIQUES AS A RESULT OF CONTINUAL EXPOSURE.

AN EXCELLENT WAY TO ACQUIRE THE NECESSARY BACKGROUND SKILLS FOR A CIVILIAN CAREER IN THIS FIELD IS TO OBTAIN TRAINING AND EXPERIENCE IN THE RELATED NAVY OCCUPATIONS.

EXAMPLES OF CIVILIAN JOBS:	RELATED NAVY OCCUPATIONS:
**********	**********

CORRESPONDENCE CLERK	PERSONNELMAN (PN) LEGALMAN (LN) YEOMAN (YN)
SPECIAL-CERTIFICATE DICTATOR	PERSONNELMAN (PN) LEGALMAN (LN) YEOMAN (YN)
TOWN CLERK	PERSONNELMAN (PN) LEGALMAN (LN) YEOMAN (YN)

****

WORKER TRAIT GROUP: INFORMATION GATHERING, DISPENSING, VERIFYING AND RELATED WORK
(DOT VOL. II PAGE 258)

WORK PERFORMED: WORK ACTIVITIES IN THIS GROUP PRIMARILY INVOLVE COLLECTING INFORMATION, CHECKING IT FOR ACCURACY OR CONSISTENCY, AND/OR RELATING OR OTHERWISE DISPENSING IT TO INTERESTED PARTIES. CARRYING OUT PRESCRIBED ACTIONS IN RELATION TO THE INFORMATION IS OCCASIONALLY INVOLVED. THE INFORMATION IS USUALLY OBTAINED OR VERIFIED BY PERSONAL OR TELEPHONE INTERVIEW, OR BY OBSERVATION, AND DISPENSED IN RESPONSE TO OR IN ANTICIPATION OF DIRECT OR INDIRECT INQUIRES. THE PERFORMANCE OF CLERICAL TASKS SUCH AS RECORDKEEPING, IS FREQUENTLY INVOLVED.

TRAINING AND METHODS OF ENTRY: A HIGH SCHOOL EDUCATION WITH EMPHASIS ON COMMERCIAL COURSES IS USUSALLY SUFFICIENT QUALIFICATION FOR ENTRY INTO THIS WORK. SOME EMPLOYERS PERFER APPLICANTS WITH SOME COLLEGE OR NIGHT SCHOOL EDUCATION.

TRAINING IS THE RELATED NAVY OCCUPATIONS PROVIDES AND EXCELLENT WAY TO GET STARTED IN ANY OF THE CIVILIAN JOBS IN THIS FIELD.

760

EXAMPLES OF CIVILIAN JOBS:	RELATED NAVY OCCUPATIONS:
**********	**********
PERSONNEL CLERK	PERSONNESMAN (PN) ENLISTED PERSONNEL ADMINISTRATIONMAN (YN-2526)
RELIGIOUS-AFFAIRS CLERK	CHAPLIAN'S CLERK (YN-2525)
PROCUREMENT CLERK	INDEPENDENT DUTY STOREKEEPER (SK-2015)

****

WORKER TRAIT GROUP: CASHIERING (DRUG STORES, THEATERS, RESTRAURANTS, AND RELATED ESTABLISHMENTS)
(DOT VOL. II PAGE 269)

WORK PERFORMED: WORK ACTIVITIES IN THIS GROUP PRIMARILY INVOLVE PAYING AND RECEIVING MONEY OR ITS EQUIVALENT AND KEEPING RECORDS OF RECEIPTS AND EXPENDITURES. TYPICALLY, WORKERS ARE EMPLOYED BEHIND CASH REGISTERS IN SUCH ESTABLISHMENTS AS DRUG STORES, THEATERS, RESTAURANTS, AND AMUSEMENT PARKS, AND ARE ENGAGED IN COLLECTING MONEY, MAKING CHANGE, CASHING CHECKS, RECORDING DETAILS OF TRANSACTIONS, VERIFYING CASH ON HAND AGAINST BILLS AND RECEIPTS, BALANCING DAILY ACCOUNTS, AND GIVING INFORMATIONS TO PATRONS.

831 832

TRAINING AND METHODS OF ENTRY: A HIGH SCHOOL EDUCATION IS GENERALLY THE MINIMUM REQUIREMENT FOR EMPLOYMENT IN THIS KIND OF WORK. COURSES IN BUSINESS ARTHMETIC, BOOKKEEPING, AND PUBLIC SPEAKING SHOULD BE INCLUDED IN THE CURRICULUM.

SOME PRIVATE BUSINESSES AND SCHOOLS HAVE SPECIAL PROGRAMS IN CASHIER TRAINING.

MANY ESTABLISHMENTS OFFER ON-THE-JOB TRAINING.

EXAMPLES OF CIVILIAN JOBS:	RELATED NAVY OCCUPATIONS:
CASHIER	DISBURSING CLERK (DK)
TICKET SELLER	POSTAL CLERK (PC)
CARUROOM ATTENDANT	SHIP'S SERVICEMAN (SH)

WORKER TRAIT GROUP: CLASSIFYING, FILING, AND RELATED WORK
(DOT VOL. II PAGE 276)

WORK PERFORMED: WORK ACTIVITIES IN THIS GROUP PRIMARILY INVOLVE GATHERING TOGETHER, CLASSIFYING, SORTING, AND/OR FILING CORRESPONDENCE, ACCOUNT RECORDS, BUSINESS FORMS, AND RELATED DATA. THE PREPARATION OF REPORTS AND KEEPING OF RECORDS IS FREQUENTLY INVOLVED.

TRAINING AND METHODS OF ENTRY: GRADUATION FROM HIGH SCHOOL WITH SUCCESSFUL COMPLETION OF COMMERCIAL COURSES AND DEMONSTRATED FAMILIARITY WITH ELEMENTRY CLERICAL SKILLS WILL USUALLY QUALIFY AN INDIVIDUAL FOR ENTRY INTO THIS WORK. MANY EMPLOYERS ADMINISTER TEST DESIGNED TO SHOW THE APPLICANT'S SUITABILITY FOR EMPLOYMENT OF THIS NATURE.

A PERIOD OF ON-THE-JOB TRAINING RANGING UP TO 3 MONTHS IS USUALLY SUFFICIENT TO FAMILIARIZE ENTRY WORKERS WITH CLERICAL PROCEDURES AND FORMS USED IN SPECIFIC ESTABLISHMENTS.

EXAMPLES OF CIVILIAN JOBS:	RELATED NAVY OCCUPATIONS:
CODING CLERK	DATA PROCESSING TECHNICIAN (DP)
	COMPUTER PROGRAMER/SYSTEM PROGRAMER (DP 2741-6)
PROGRAMER DETAIL	COMPUTER PROGRAMER (DP 7426)
MEDICAL-RECORD LIBRARION	MEDICAL ADMINISTRATIVE TECHNINIAN (HM 8442)
	FIELD DENTAL TECHNICIAN (DT 8707)
FINGERPRINT CLASSIFIER	

WORKER TRAIT GROUP: COMPUTING AND RELATED RECORDING
(DOT VOL. II PAGE 280)

833

834

WORK PERFORMED: WORK ACTIVITIES IN THIS GROUP PRIMARILY INVOLVE PERFORMING ARITHMETIC COMPUTATIONS AND PREPARING NUMERICAL RECORDS WITH THE AID OF TYPEWRITERS AND ADDING, BILLING, BOOKKEEPING, AND CALCULATING MACHINES. TYPICAL ACTIVITIES ARE COMPUTING WAGES, INTEREST, AND PRODUCTION COSTS; TOTALING BANK DEPOSIT SLIPS, AND VERIFYING COMPUTATIONS RECORDED IN ACCOUNTS.

TRAINING AND METHODS OF ENTRY: GRADUATION FROM HIGH SCHOOL AND BUSINESS SCHOOL WITH AVERAGE OR SUPERIOR GRADES IN ARITHMETIC FREQUENTLY IS THE MINIMUM REQUIREMENT FOR ENTRY INTO THIS TYPE OF WORK.

FOR SOME KINDS OF WORK SPECIAL TRAINING ON A PARTICULAR MACHINE IS REQUIRED. FOR MANY BEGINNING POSITIONS, HOWEVER, A GENERAL KNOWEDGE OF THE WORK AND THE EQUIPMENT INVOLVED USUALLY IS REGARDED AS SUFFICIENT PREPARATION. OFTEN AN EMPLOYER WILL GIVE A NEW EMPLOYEE INSTRUCTION AND ON-THE-JOB TRAINING, RANGING ANYWHERE FROM A FEW DAYS TO SEVERAL MONTHS, DEPENDING CHIEFLY ON THE TYPE OF WORK OR THE KIND OF MACHINE INVOLVED.

A GOOD WAY TO GET A JOB IN THIS FIELD IS TO TRAIN FOR IT IN THE RELATED NAVY OCCUPATION.

EXAMPLES OF CIVILIAN JOBS:	RELATED NAVY OCCUPATIONS:
PAYROLL CLERK	DISPURSING CLERK (DK)
AIRCRAFT LOG CLERK	AVIATION MAINTENANCE ADMINISTRATION MAN (AZ)
ACCOUNTING CLERK	STOREKEEPER (SK) AVIATION STOREKEEPER (AK)
BOOKKEEPER	DISBURSING CLERK (DK)

762

SUMMARY

A NUMBER OF CIVILIAN AND NAVY OCCUPATIONAL POSSIBILITIES HAVE BEEN PRESENTED TO YOU. THEY MAY OR MAY NOT BE IN LINE WITH WHAT YOU HAD IN MIND AS PROBABLE CAREER FIELDS. DEPENDING UPON YOUR AGREEMENT WITH THE LISTED OCCUPATIONS YOU HAVE SEVERAL OPTIONS WHICH YOU CAN FOLLOW. THERE ARE OUTLINED BELOW:

(1) YOU ARE SATISFIED WITH THE ABOVE JOBS LISTED, THAT IS, YOU CAN MAKE CAREER DECISIONS BASED UPON THE INFORMATION PROVIDED.

(2) YOU CAN MAKE CAREER DECISIONS BASED ON THE ABOVE INFORMATION BUT WOULD LIKE ADDITIONAL INFORMATION ABOUT THE JOBS: SEE YOUR SCHOOL GUIDANCE COUNSELOR OR NAVY RECRUITER, AS APPLICABLE.

(3) YOU ARE NOT SATISFIED WITH THE OUTPUT IN THAT THE CIVILIAN JOBS DO NOT CORRESPOND WITH YOUR OCCUPATIONAL PLANS. IN THIS CASE, IF YOU DESIRE FURTHER COUNSELING YOU MAY TAKE ONE OR ALL OF THE FOLLOWING ACTIONS:

(A) RETAKE THE ASVAB - THE JOBS WERE SELECTED ACCORDING TO JOB APTITUDE REQUIREMENTS AND YOUR ASVAB SCORES. IF YOU FEEL THAT THE ASVAB SCORES ARE NOT A TRUE REFLECTION OF YOUR APTITUDE AREAS YOU CAN ENTER NEW ASVAB SCORES IN THE COMPUTER SYSTEM AT ANOTHER COUNSELING SESSION. TO RETAKE THE ASVAB, CONTACT A NAVY RECRUITER.

835 (B) YOU MAY DESIRE TO CHANGE YOUR EDUCATIONAL LEVEL (QUESTION 1, PAGE 3, OF YOUR BOOKLET). THE JOBS SELECTED AND PRESENTED ON SECTION I OF THIS PRINTOUT WERE BASED, IN PART, UPON THE MINIMUM EDUCATIONAL LEVEL REQUIRED FOR ENTRY INTO THESE JOBS. THEREFORE THE HIGHER LEVEL OF EDUCATION YOU ANTICIPATE COMPLETING THE GREATER NUMBER OF OCCUPATIONS THE COMPUTER WILL SELECT FOR YOU. HOWEVER YOU SHOULD BE REALISTIC. THERE IS NO POINT IN INDICATION THAT YOU PLAN TO COMPLETE 4 YEARS OF COLLEGE AND HAVE THE COMPUTER LIST ALL THE OCCUPATIONS THAT 836

REQUIRE COLLEGE EDUCATION WHEN IN FACT YOU ARE CERTAIN THE FURTHEST YOU ANTICIPATE COMPLETING IS A TECHNICAL SCHOOL OR APPRENTICESHIP PROGRAM.

(C) YOU ALSO HAVE THE OPTION OF CHANGING YOUR INTEREST CATEGORIES. IN ADDITION TO SELECTING THE OCCUPATIONS BASED UPON YOUR APTITUDES AND EDUCATIONAL PLANS THE JOBS PRESENTED IN SECTION I WERE IDENTIFIED ACCORDING TO YOUR OCCUPATIONAL INTEREST AS SHOWN IN QUESTION 2 PAGE 3 OF YOUR WORK BOOKLET. IF YOU ARE NOT REASONABLE CERTAIN OF YOUR SELECTIONS YOU MAY DESIRE TO TAKE AN INTEREST INVENTORY. MOST INTEREST INVENTORIES WILL TELL YOU HOW HIGH OR LOW YOUR INTEREST ARE IN RELATION TO PEOPLE WHO ARE WORKING IN VARIOUS OCCUPATIONS. IF YOU DESIRE THIS TYPE OF HELP SEE YOUR COUNSELOR OR A NAVY RECRUITER.

763

Rafacz, B. A. The selection of an occupational counseling computer network for the navy recruiting environment. A presentation made to the 18th Military Testing Association Conference. Pensacola, Florida October, 1976.

Yellen, T. M. I. A computerized guidance and occupational system for civilian and related navy occupations. Navy Personnel Research and Development Center, San Diego. In progress.

The Selection of an Occupational Counseling Computer Network System for the Navy Recruiting Environment*

by

Bernard A. Rafacz

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

Problem

The problem within Navy recruiting has always been to enlist a sufficient number of recruits to satisfy present and projected Navy manpower requirements. To minimize A-school attrition and requests for reassignment, the Navy recruiter and classifier attempt to enlist and seek assignments for recruit applicants that result in a satisfying Navy career. Furthermore there is a need to "counsel" applicants so that the subsequent pool of recruits satisfy the recruiting goals of recruiters, districts, and Areas within the Recruiting Command. Needless to say there is a tenuous decision process (a 2-person game, if you will) between the recruit applicant, and his desire to obtain that assignment or committment from the Navy which best match his present aspirations, and the Navy recruiting system, which hopes to enlist the applicant where the need is the greatest.

To assist the Navy manpower planner in the fundamental problem of enlisting a sufficient pool of recruits there are tools and techniques that could be utilized but for various reasons are not employed. There are three possible explanations for this. The first and most obvious explanation is that many of these techniques are unknown to the recruiter classifer or manpower planner. The second explanation is, even if the recruiter classifier or manpower planner were knowledgeable in what could be utilized, there is not an appropriate delivery system upon which to implement the techniques. The third explanation is that some techniques are yet in a research and development phase and are only now being evaluated for future utilization. In brief, the problem is that of bridging the gap between the research and applied areas of the Navy recruiting effort.

The issue to be addressed within this report is the identification of a delivery system that could eventually be utilized within the Navy

*The opinions expressed are solely those of the author and do not necessarily reflect official Department of the Navy policy. Reproduction in whole or in part is permitted for any purpose of the United States Government.

recruiting environment to assist in the fundamental problem of counseling and recruiting applicants for Navy Careers while simultaneously addressing the data management requirements of Navy Recruiting manpower planners from report generation to accessions control.

Scope

The scope of this report refers to all of the available, or anticipated, computer network systems which could qualify as delivery systems. Of course, the choice of a system is dependent upon the anticipated functions of the system once implemented. In some instances the functions already exist as part of other military or civilian manpower systems, while some functions are as yet in a developmental research mode but nonetheless deserve attention as possible application areas. It was decided that the delivery system must entertain, as a minimum, the application areas of data base management (DBM), computerized counseling, and the sequential assignment of recruit applicants. Briefly, data base management refers to the collecting, maintaining, and analyzing of relevant recruiting data. This implies applications in such diverse areas as computerized testing, recruiting goal generation procedures, and the producing of summary reports for manpower planners. Computerized counseling by the delivery system is the capability of the system to maintain a number of interactive dialogues with a recruit applicant, monitor his career decision processes, and produce career related information either on-line or via a summary report. The sequential assignment of applicants refers to the capability of the system to entertain applicant preferences and interests while attempting to maximize the value of some assignment criterion function. This is in contrast to a first come first serve policy which does not weigh an applicant with the quality of the entire applicant pool nor attempt to optimize an assignment criterion function.

Background

Military Assignment/Reservation Systems

Army The Army has been developing project REQUEST. This project involves an interactive real-time sharing system utilizing the INFONET communications network. Software for the system was developed by Systems Automation Corporation. The project has terminals located in the Guidance Counselor Offices of each of the 74 Armed Forces Examining and Entrance Stations (AFEES). In addition to the terminals at the AFEES, the Army has assigned some to mobile recruiting teams and increased the number of terminals to over 200. The GSA has contracted the Computer Sciences Corporation (CSC) INFONET system to provide nationwide linkages from AFEES to a UNIVAC 1108 computer complex near Los Angeles.

REQUEST presently includes all enlistments and allocation to all Army training spaces. It is basically an airline reservation system and all reservations for school training spaces are made on a first come first serve basis. The counselor at AFEES is the only interface between

the Army applicant and the REQUEST system. After an applicant-counselor interview of approximately 30-40 minutes the counselor enters on the REQUEST terminal all information on a prospect's physical profile, test scores, desired entry dates and desired occupational area. REQUEST matches this information against requirements for various occupational areas and the available quotas. Once the search is completed the best matches are displayed and the counselor and applicant decide on an occupational area. The processing transaction time on the terminal consumes about 10 minutes of wall-clock time.

Air Force The Air Force has been conducting research on an Advanced Personnel Procurement System (APPS) that addresses alternatives to its Accessions Control Center (ACC). Hawkins, Crow, and Haltman (1974) report on six different configurations of an APPS for the Air Force. The most advanced configuration of computer software and hardware evaluated consists of teletype terminals operating at 110 baud located in each of the Air Force's 997 recruiting offices. By a person-job match algorithm a computer system will examine all job vacancies available and select those training slots or aptitude areas in which the individual is best qualified to train, and match the prospect's abilities and interests. The most sophisticated configuration evaluated in the report is essentially a configuration that interacts with a recruiter. The other five configurations include the location of terminals in recruiting sectors, detachments, or groups, at AFEES, or upgrading the presently utilized ACC.

As of this date the Air Force is in the process of implementing that configuration of APPS involving the location of terminals at AFEES. As with the REQUEST system, a member of the liaison staff at AFEES would access a terminal after the applicant has completed mental and physical testing. Essentially this configuration transfers job counseling from field recruiters to the liaison staff at the AFEES station. Necessarily success of this configuration and the REQUEST design, is dependent upon the willingness and the feasibility of permitting prospects to make up their minds about a job while at AFEES. Otherwise there will be a continuation of recruiter phone contact for assignment of prospects. Furthermore it is not clear the extent to which it is possible for a recruiter to divorce himself of the being a counselor and become merely a referral agent.

Navy Most recently the Navy has been developing, through a contract effort with System Automation Corporation (SAC), project PRIDE (Personalized Recruiting for Immediate and Delayed Enlistment). This effort has essentially replaced the manual Recruit Allocation Control System (RACS) which heretofore accounted for the processing of accessions with guaranteed Class A-school seats. Project PRIDE at present consists of terminal operators at Navy Recruiting Command Headquarters receiving, via commercial phone lines, requests for school seats from field recruiter/classifiers. Software for the PRIDE system was developed by SAC as a Navy version of the Army's REQUEST system. Therefore, as in the REQUEST system, once an operator receives a request for a school

seat, the operator enters the appropriate information and the system responds to the terminal operator of the opportunities which are presently available to the applicant. The operator relays this information to the recruiter/classifier and a decision is made as to which seat the applicant will accept. Finally the operator reserves the seat for the applicant and records a profile of data on the recruit.

In the next phase of PRIDE it is expected that the terminals will be colocated with AFEES in the field and physically located at a Navy A-station or a district headquarters office near the AFEES. All such locations will include a classifier. This design is prompted by the role of the classifier as being the counselor in terms of Navy career areas, and the recruiter's role as that of "selling the Navy." As in the Army's REQUEST system, or the Air Force's APPS, it is expected that the applicant will decide upon a job or career area while at AFEES.

Project CONTRACT For the past several years NPRDC has been conducting the research effort identified by the acronym CONTRACT (Computerized Navy Techniques for Recruiting, Assignment, Counseling, and Testing). Besides developing techniques within the areas of DBM and computerized counseling, testing, and assignment, the project is responsible for identifying the vehicle for implementing such techniques within the Navy Recruiting environment. Rafacz (1975) reports on the research efforts of project CONTRACT.

Distributed Computing

Probably the most prominent development in computing thus far in the seventies has been the emphasis and associated technological developments in distributed computing. The concept of distributed computing within a multiple processor computer configuration was studied in the sixties, with the result of a number of successful multiple processor computer systems, such as the Burroughs 7700. In these processors, the computing load is distributed among two or more processors. An extension of this concept was to configure computing networks in which the processors are housed in different cabinets, with extremely high speed connections. They are then located in different buildings, different parts of town, different cities, and communicate to each other over telephone lines. Attempts are being made to configure processors appropriate to each function which must be performed in a data processing system. For instance, source data automation is concerned with the development of specialized intelligent terminals to format CRT's for forms entry, perform editing of keyed data, and prepare data for later transmission to another processor. Communications processors are being employed to solely perform the data communications functions of computing systems, thereby freeing another processor from performing these tasks.

APPROACH

As previously mentioned, the purpose of this report is to outline the design for a delivery system that is capable of entertaining some of the operational functions of a system such as Navy Recruiting, not as it presently exists, but rather as it may be in the foreseeable future. Necessarily it is required to anticipate those functions which would be a part of the system at that time. Justification for the selected functions will emanate mainly from what is already and operationally a part of other systems which, in some respects, resemble the Navy Recruiting System. Further justification can be derived on the basis of work which is presently being developed and evaluated within the research community, and in particular, that of NPRDC.

Anticipated Functions

This report will address three general functional areas which are to at least a small degree already a part of the Navy Recruiting System, although not necessarily as presently interpreted by that system. The three areas of concern have been identified as Data Base Management (DBM) techniques, Computerized Counseling procedures, and the Sequential Assignment of recruit applicants. These functions were selected because it is clear that they would enhance the recruiting efforts of the Navy Recruiting Command by replacing the present "manual" procedures employed to satisfy an operational requirement. It is hypothesized that an enhancement of present procedures would result in a savings of personnel manhours and provide Navy personnel with additional tools and techniques to supplement their present efforts. Specific payoffs have already been realized in some areas while additional payoffs are obvious once system implementation becomes a reality. Furthermore it is possible to point to some civilian and military manpower systems which have already implemented the areas to some degree and are realizing manpower savings besides added system capability.

Data Base Management Techniques

In general, data base management (DBM) will refer to the capability of a manpower planner to create, maintain, and sort a file (e.g., a data set on accessions; each accession in the file is represented by a complete collection of information). Once files of interest have been created, a manpower planner can request that computations be performed, process queries (requests for information) on specific individuals in the file, and produce reports on the entire file or subsets of the file. This is particularly relevant to reports such as the Weekly/Monthly Telephone Report required by each Area to submit to CNRC and for each District to submit to the Area Commanders. However, over and above required reports, each district and Area Headquarters generate unstandardized reports that are to be utilized to satisfy the requirements of

their specific operational environment. These reports are generally an accounting on accessions but could also include reports on vehicle utilization, or recruiter effectiveness. Furthermore, there is also the need for timeliness of report generation. Present manual procedures greatly hinder (or deny) the generation of some reports simply because too many manhours of effort would be required for their generation. A computerized procedure is required to maintain files of data, determine a manpower planner's needs, and then to produce the required report within minutes of a request.

Desirable Characteristics of a Navy Recruiting DBM System There is not too much that could at present be inferred which would mandate unique requirements of a DBM system for Navy Recruiting. In fact, Navy Recruiting is very much like any other large business organization in that certain supply, fiscal, and personnel data is collected and it is required that certain reports be generated on a periodic basis (daily, weekly, monthly, quarterly, etc.). Furthermore it is reasonable to expect that Navy manpower planners would also desire the capability of generating ad hoc reports. Therefore certain specifications were included in the system design which would be of interest to most any DBM shopper, to include: support for both relational, network-like and hierarchical data structures, file load processors, deadlock protection at the record level, file, record, and field-level security, support for multiple files (and multiple "data bases"), inverted file capability, support for both sorted (indexed-sequential type) and unsorted (random) files. Furthermore, it was felt that the DBM system should support certain additional features to act as a safeguard against expanded utilization of the system once implemented. Clearly the availability of a DBM system in an environment that essentially operated heretofore via manual procedures (below the Headquarters level) would invent new and innovative uses for the DBM system. Therefore a certain buffer must be built into the specifications for a DBM system for the recruiting environment. In view of these observations the following represents a list of those requirements which were believed to be necessary for the DBM software of the final computerized monitoring system:

a. Is reentrant.

b. Allows for concurrent batch and on-line access to the DBMS and its data.

c. Automatically goes to end of job and relinquishes its memory resources when application no longer accesses the data base.

d. Centrally manages data buffers itself rather than forcing applications to do so.

e. Allows validation of incoming transaction data.

f. Permits multiple access techniques for same data items without multiple redundant files.

g. Has inverted file capability.

h. Allows recovery from record dead locks for batch programs.

i. Can recover from fatal errors to the last transaction if necessary.

Computer Assisted Counseling

The concept of computerized counseling within the dimensions of this report is meant the capability of a computer system to maintain a real-time interactive dialogue (i.e. and intelligent discourse) with a Navy recruit applicant. Recruiter or classifier assistance during the course of the dialogue is minimal. The ultimate purpose of the dialogue(s) is to provide the applicant with personalized complete, uniform and unbiased information concerning Navy job occupations and the Navy in general. It should not be inferred that such dialogues would replace the counseling role of recruiters/classifiers, rather it is meant to supplement their efforts in this regard. A wide range of occupational information from numerous sources and continually updated text file information should characterize the counseling system. The payoff, of course, is that the computer can perform the job skill matching of an applicant (based upon the applicant profile of data collected via computerized testing) over a much wider range of job opportunities and dimensions than what could be normally possible. Such a more thorough career decision process would permit the applicant to adequately explore the appropriateness and correlation among his personnel desires and available opportunities. Necessarily computer assisted career counseling will impact in a positive way on the applicant career aspirations, and subsequently on overall Navy manpower requirements and organizational effectiveness. See Yellen (1976) for a more detailed account of a computerized counseling methodology.

Characteristics of a Navy Recruiting Computer Assisted Counseling System There are several computer assisted counseling systems in existence and the reader is referred to the reports of Impelliterri (1968) and Katz, Chapman, and Godwin (1972) as examples of such systems. It is not the intention of this report to elaborate on each as to its more desirable features, some of which may be included in a computer assisted counseling system for Navy recruit applicants. Rather, a review of present computerized counseling systems have yielded valuable information as to what could possibly be accomplished given the availability of a computer system within the Navy Recruiting environment. Furthermore, there are some functions that could be performed in a computer assisted counseling mode which are not a part of present systems, but which it is felt should be a part of a Navy Recruiting counseling system. The following list represents a scenario of functions and characteristics which are considered to be required of a system that could be utilized within the Navy Recruiting environment for the career counseling of Navy recruit applicants.

1. Applicant-system Interface. Refers to the availability of a device by which information can be received (transmitted) by applicants from (to) the computer system.

2. Applicant Reading Ability Monitoring. In an environment such as Navy Recruiting there are many types of individuals seeking career counseling with varying reading abilities. The system must be able to detect (via a short reading test), or upon being cued by the recruiter/classifier, the general reading ability of the applicant.

3. Level of Career Counseling Adaptibility. The logic of the computer counseling dialogue must be geared to the particular level at which the applicant is seeking counseling. An applicant that is obviously not sure of his career goals or opportunities must be distinguished from one who has already determined his career goal but is seeking more detailed information about a specific career area or training opportunities within the area.

4. Retrieval of Applicant Input Data. If an applicant has been processed through a counseling session in a prior dialogue, the system must be able to retrieve any career related information (from any prior dialogue) for the present dialogue without the need for the applicant to resubmit the necessary information.

5. Detection of Inconsistencies in Logic. The logic of the counseling dialogues must be capable of identifying inconsistencies in the personal desires (e.g., aspirations, interests, and job preferences) of the recruit applicant with respect to his mental and physical capabilities.

6. Modification of Applicant Profile Data. The counseling system will allow the applicant to modify his input data (except that which is fixed for a given applicant, such as aptitude scores) at an appropriate time during the course of a dialogue.

7. Screening of Applicant Profile Data with Opportunities. Within some dialogue, the counseling system must be capable of matching an applicant's profile with a set of civilian occupational areas and corresponding Navy Programs for which the applicant qualifies.

8. Generate Navy Career Path Flow. Once the applicant has identified desirable (or willing to enlist) Navy occupational areas, it would be possible to provide the applicant with a simulated career path of selected Navy careers.

9. Produce Summary Reports. While operating in a real-time interactive mode, the counseling system will be able to produce short summary reports of any civilian or Navy job opportunity available to the

applicant, particularly those Navy Training schools which best match an applicant's profile.

10. Availability. The computer system and associated counseling logic must in all instances be geared to the end user's unique needs.

The Optimal Assignment of Recruit Applicants

In the present military recruiting climate a recruit applicant is offered a training program (job area or school) if certain conditions are met; these being that the applicant is mentally, physically, and morally qualified. In addition the applicant must be willing to be enlisted into that program. Furthermore, the quotas must not be filled for the school for the time at which he is applying. This policy is reminiscent of the first come first served assignment strategy of the REQUEST system operating at AFEES. The problem with this policy is that it restricts offering assignments to more qualified applicants who desire a particular school assignment when the school quota has already been filled by applicants who have received their assignments earlier in a recruiting time frame.

Traditionally, within the Navy recruiting environment, the optimal assignment of recruits to Navy training schools is made while the recruits are in basic training. This optimal assignment technique, called COMPASS (Computer Assisted Assignment System) maximizes the value of recruit assignments with respect to three dimensions; recruit preference, travel cost to the training station from the RTC, and the recruit's aptitude test scores. However, as recruiting policies are currently being directed towards an increased use of assignment commitment prior to actual enlistment, the utility of the present optimization procedure is substantially diminished[1]. There is a need to implement a procedure for anticipating highly qualified applicants and to reserve school seats for their arrival. In short, applicants would be required to compete with the entire applicant pool; not only those which have already arrived into the system, but those yet to arrive. The sequential assignment strategy has been identified as the procedure which would enable a manpower planner to offer only "optimal" school seats to applicants, with due regard for the quality and aspirations of the entire applicant pool. Further, that strategy is required to better fill required quotas than the presently utilized first come first served strategy.

Sequential Assignment Procedures NPRDC researchers have developed several sequential assignment procedures which hold sufficient promise for application within the Navy's recruiting environment. These techniques require data on a recruit or applicant population that has already enlisted at a prior point in time. This prior population resembles the anticipated applicant pool with respect to an assignment criterion function. The reader is referred to the reports of Horst and Sorenson

[1] At least 50% of the present recruit population have been receiving school guarantees of some type.

(1976) and Rafacz and Halstead (1976) for more detailed information concerning various approaches to the problem. Yet another significant effort is that of Ward (1959). Ward's sequential assignment technique, together with a sophisticated assignment payoff function will be a part of the Air Forces APDS-PROMIS system to be implemented in Fall, 1976.

Functional Requirements of a Sequential Assignment Procedure The sequential assignment procedures surveyed for purposes of this report indicate that the actual algorithms employed to accomplish the assignment process are relatively simple, from a computer processing standpoint. However it is not feasible, even for the easiest technique considered, to expect the technique to be applied manually. Furthermore, parameters utilized in some techniques require updating of a prior population to reflect the anticipated applicant pool for the time period under consideration. The most difficult part of the process is the requirement that the record of assignment reservations be updated on an almost real time basis; i.e. the assignment system must have an almost airlines reservation capability to any level below the AFEES and Area and Command Headquarters strata; with real time reservations at those levels. The following list represents those minimal requirements of the sequential assignment/reservation system of project CONTRACT.

1. The sequential assignment algorithm employed must weight an applicant with the quality of the entire applicant pool with respect to a predetermined assignment payoff function. In application the technique must, as a minimum, be an improvement of a first come first served policy with respect to the payoff function and closely approximate an optimal assignment solution when applicant preferences are disregarded. Furthermore, the technique must fill assigned quotas to an extent at least as well as a first come first served policy.

2. The assignment/reservation capability must be available to every user at every terminal within the computerized counseling system which has access to the DBM system. If the terminal is at an application strata below AFEES or Area Headquarters, the reservations must be no more that 30 minutes behind real time.

3. The system must be able to provide an accounting of every accession for the present recruiting time frame into any possible enlisted program or enlistment type with respect to recruiting goals at the Command, Area, and district levels [e.g., overall Navy, Area and district recruiting goals, RTC Shipping Schedules (up to six months in advance), Minority Recruiting Goals, Seafarer/Airman Programs, Reserve programs, and Advanced Technical Field Quotas].

4. The reservation capability of the system must have all the capabilities of an airlines reservation system; the real-time aspect of the system can be restricted to the extent mentioned in 2. above. This includes the capability of the system to process queries and updates upon a centralized accession file; recording as a minimum the actual location of pertinent data on accessions.

5. As opposed to accessions, the system must record pertinent data on applicants, at least those which were processed to the point that one counseling dialogue was completed and/or the ASVAB (or similar aptitude test) was administered. Summary reports on applicants must be available within 72 hours of a request by Command, Area, and district manpower planners.

Development of Evaluation Criteria for Possible System Configurations

The system functions described in some detail in previous sections should imply certain hardware and software specification of the eventual monitoring system. Of course, it is no easy matter to translate those functions into specific hardware and software requirements. To accomplish this necessary task it was decided that certain evaluation criteria for hardware and software be established before the actual selection process. Likewise certain "considerations" involving the system functions must be considered in order to assist in the development of such criteria. This facet of the developmental effort was refered to as applications assessment.

Computing application areas encompass a wide diversity of subjects; data base maintenance of financial and other management information, counseling of prospective military personnel on both military and civilian job opportunities, school seat status querying, assignment of personnel to appropriate military occupational specialities, and possible testing of applicants in the field. Concurrently with these requirements is the fact that due consideration must be given to the organizational structure of the Navy Recruiting Command. At present, NRC is a hierarchcal organization consisting of a central headquarters, six Area offices, 43 District offices and over 1400 recruiting stations. The important consideration is that each of the diverse applications has its own set of computing resource utilization characteristics within the levels of the organization. Further, within each application, there are time-dependent computing requirements. For instance, reservation requests for highly competitive school seats must be serviced quickly. On the other hand, retrieval of data by the command headquarters from remotely located districts could be delayed by as much as a day.

NETWORK DESIGN

The proposed network for the Recruiting Command District, Area and National Headquarters stations has been designed for maximum reliability, capability and expandability. Every attempt has been made to give the Command capabilities which are available only through a dedicated private, Navy-owned network of computers. The general philosophy has been that of a network which is distributed over a wide geographical area but which provides centralized control over testing of equipment and circuits at every level in the network.

Centralized Vs Distributed Computing

For purposes of this report, centralized systems are characterized by one or several multi-million dollar processors located at one central geographical location. A distributed computing network is described as the availability of many small processors (minicomputers) typically located as near the end-user as possible. Job requests by a user to a centralized system must be via telephone lines from his remote location, normally at a 300 baud data transmission rate. In a distributed network, job requests can normally be entertained by the user's local processors at from 1200 to 9600 baud. Furthermore, minicomputers can be assigned to perform specific tasks via inter-processor communication procedures. An example of a centralized computer network system is given by the INFONET communications system and the UNIVAC 1108 computer located in Los Angeles and presently being utilized by the Army for the REQUEST project. An example of a distributed processing system is given by Security Pacific Bank. Security Pacific has an IBM 370 computer as the central processor with communications links to General Automation minicomputers located in branch offices to perform most of the communications protocol.

The functionally distributed computer network system was selected as the most appropriate to serve as the vehicle to implement research developed tools within the Navy Recruiting environment. Underwood and Rafacz (1976) describe in more detail the justification for this selection. The following outline is provided to serve as a partial rationale for purposes of this report:

1. Reliability. In a centralized system, when the central computer is incapacitated, every user is also incapacitated. In a distributed processing system only users assigned to the incapacitated minicomputer are without computer access.

2. Telecommunications Charges. Every user must assume a telecommunications charge from his remote location to the central computer site for centralized systems. Users may have no telecommunications charge for processors at their location in a distributed network, or an "across the town" telephone call charge at worse.

3. System Expandibility. A centralized system requires a very large initial investment. Distributed systems can be developed as functions become better known and funds become available.

Distributed Network Components

The chain of command in the Recruiting Command is followed, generally speaking, in determining the allocation of computing resources. A large multiple-processor minicomputer system will be located at the Great Lakes naval installation. This is a strategic location so far as minimization of communications costs are concerned. The function of this computer system will be to monitor the network. It will keep track of the load on the aggregate network, in terms of computing resource utilization and

in terms of communications costs. It will be the clearing house for ad hoc and scheduling reports required for the Recruiting Command headquarters in Arlington, VA. It will also act as the primary file for school seat reservations in the network. Other processors, located at the Area offices, will routinely and periodically during the day interrogate this file, update it by making requested reservations, and in turn receive information on transactions that have been completed since the last call. In this way, each Area processor will redundantly possess its own reservations file which can be accessed by the Area and its district office processors for near-real time information on school seat availability. Also, the monitoring multiprocessor will possess a 2780 protocol communications line to the Recruiting Commands IBM 360 in Arlington, VA. Special purpose batch jobs and information relative to files maintained by headquarters will be sent over this 2000 bps line.

CRT's will be located in Arlington and will communicate over a multiplexed line to the monitoring multiprocessor. In this way, Headquarters may communicate with not only the Great Lakes computer configuration, but with any processor in the network, using a message control environment in which messages are exchanged along with reservations data among calling Area processors, which in turn communicate with their District processors in real time. Headquarters can propagate changes in goaling on short notice, using this kind of message control environment and NPRDC-provided goal generation software which is available on all Area processors.

Area Processors

The Area processors may consist of a Burroughs B1700 line, or equivalent, system, similar to the one which is in procurement as a prototype by NPRDC. Area offices will have hard copy capability in terms of a 250 lines-per-minute printer and hard copy devices associated with CRT's. Area processors will possess accounting-type intelligent terminals in addition to teletypewriter CRT's; these terminals (such as the Burroughs AE 500 or the General Automation) will facilitate the use of the Area computers for maintaining accounting records by warrant officers and civilian and military clerical personnel. These terminals possess some programmable capabilities themselves, apart from the Area processor. For example, they can be programmed to prompt for accounting information and to edit incoming data, batch it up and eventually channel it down a communications line to the Area processor. The capability allows for freeing up of some of the main computer's resources, and allows for greater ease of data entry for clerical personnel. Data management for the Area will be performed using specialized software that will be available on the Area processor. This processor will also automatically dial up the monitoring multiprocessor with a batch of reservations requests that have been queued up from the District computers. If possible, these reservations will be made. If not available, suggestions will be made and a reporting transaction entry made to be routed back to the requesting recruiter. Messages to be sent to Arlington and messages back from Arlington are exchanged. Any requests

for data from the Area files or from files located at District processors are satisfied at this time.

District Processors

Each District office in the Recruiting Command will have its own processor. The same kinds of data base management functions as required for the Area will be performed by the District offices. The load for Districts, especially the larger ones, is considerable. Also, counseling of recruit applicants in fixed-location (as opposed to "van stations", to be discussed shortly) stations will be performed, and again, message communications between stations and the District can be made using the data communciation lines, at least for "type-able" messages.

Mobile Units

To improve upon the current situation of the recruiting station, in which potential applicants are mostly walk-in, the proposed network would make an active attempt to get closer to the applicant pool, particularly by going to District high schools with information on job opportunities in the Navy, and providing a computerized counseling service. A roaming van concept is envisioned to provide this service. A prototype van has already been acquired by NPRDC, and some field testing will be done in FY77. The eventual network would see each van possessing its own computer configuration which did on-line counseling, some reservations requesting, general accounting, and controls some audio and visual equipment for better conveying Navy career information.

The computer configuration for the vans has been determined, although the final computer vendor has not been selected. The configuration will consist of two computers, one dedicated to performing the heavy counseling and job-matching algorithms, one dedicated to servicing the communications required for the terminals. However, the two computers will be physically located in the same cabinet, and they will share a fast fixed-head disk drive with 1 million bytes of storage capability, and a slower 10 million byte cartridge disk drive. The configuration will be expandable in order to service from 8 to 50 terminals.

Evaluation Criteria

There were many factors which were seen as relevant to the applications to be supported by the proposed distributed network. Some general guidelines are presented here, a priori considerations which directly affected the configuration hardware and software selection outcome:

1. Capability for both generalized software (such as a DBMS) and highly specific software (e.g., for string manipulation and bit manipulation) was required in order to allow for system optimization in critical areas while allowing for quickly produced less efficient code in noncritical areas.

2. The configuration selected should possess some characteristics of fourth generation hardware in order to allow for configuration continuity throughout the foreseeable project life.

3. Configuration characteristics should be such that telecommunications costs are minimized.

4. A variety of software should be available so as to allow for some matching of the appropriate language/compiler with the particular application.

5. As much off-the-shelf software as possible should be obtained so as to minimize software development costs for the project.

6. Data communications response times are assumed to be critical for all applicant-system transactions.

7. The training and labor costs associated with operators make self-loading, self-organizing configurations highly desirable.

See Underwood and Rafacz (1976) for a more detailed account of considerations involved in evaluation criteria relative to the dimensions of Job Mix Control, Operating System Capabilities, DBMS Considerations, and Data Communications. This report will include a brief discussion of configuration hardware and software with regard to the prototype minicomputer for the selected computer network.

Prototype Selection

A Burroughs B1700 system was selected as the main prototype configuration. The primary factors in our decision were the performance of the 1700 line in typical job mixes for our applications, and the presence of the most relevant features: virtual memory, automatic program reentrancy, a Network Definition Language compiler, DMS II, FORTRAN, COBOL, RPG II, User Programming Language (for string manipulation and implementation of more sophisticated data manipulation algorithms), and REPORTER. The B1700 line is expected to be the best system for handling complicated, dynamically changing job mixes, even though it was more expensive and not quite as fast for a single job. It was outstanding in the system's monitoring of resources used by applications programs, a feature essential to fine tuning of job mixes for field use, and was best in its automatic allocation of resources on an as-available and priority basis to contending jobs.

Our study of the prototype will determine in more detail the relevant performance characteristics of the B1700 systems for specific job mix environments, in order to determine the load the system can take, and thereby to configure the larger network with the appropriate number of processors. Studying the performance of such software systems as NDL and DMS II will require special techniques and a considerable investment of labor.

IMPLEMENTATION PLANS

It is clear that the implementation effort underlying project CONTRACT is probably the most crucial aspect of the entire research effort. The project is concerned with the development of computerized procedures and a monitoring system which could serve as a vehicle for researcher developed DBM, assignment and reservation, and counseling procedures. Implementation will essentially involve resolving any remaining practical utilization aspects of the proposed system.

As can well be imagined, implementation of the prototype system on a nationwide scale must occur over several phases. At present three distinct phases are viewed as necessary so that implementation on a nationwide scale can occur. Briefly these phases consist of:

Phase I: Development of a true prototype Navy recruit applicant counseling system. This effort will be concerned with developing procedures for implementing research developed counseling tools and DBM system via the prototype system. It is expected that once the system is adequately refined, researchers will have a direct line to a field application area via the prototype system. Trial runs of the system will be conducted at various levels within Navy Recruiting Area EIGHT.

Phase II: Development of a computerized counseling system for the Pacific Southwest area (includes the states of California, Arizona, Utah, and Nevada). This phase of the research effort will be primarily concerned with developing a network of computers to serve as the monitoring system for most of the recruiting stations of Area EIGHT. The prototype system will be further refined to include the testing and evaluation of more sophisticated computerized counseling dialogues. It will also be possible to collect recruit applicant data and create a data base for the research efforts involved in further refining the prototype computer system (e.g., optimal assignment). The operational aspects of the monitoring system will be identified and further evaluated together with a clearer picture of the true trade-offs of the system in terms of cost benefits and over-all effectiveness[1].

Phase III. Nationwide Implementation Effort. A cost analysis of the entire system together with specific plans for incorporating the monitoring system into then existing Navy computer systems will be completed. Refinement of the prototype will continue and revisions implemented as appropriate and feasible. It is anticipated that research efforts in this phase will involve developing procedures to enable Navy manpower planners to achieve a high level of accessibility of recruiting type data. Inherent to this effort will be the incorporation of the optimal assignment strategy (identified in the prior phases), that is expected to be of maximal benefit to the Navy and the recruit applicant.

---

[1] A report by Underwood and Rafacz (1976) explains in more detail the cost and tradeoff of various implementation phases based upon presently available data.

SUMMARY AND CONCLUSIONS

Project CONTRACT is concerned with developing computerized procedures and techniques for utilization by the Navy Recruiting Command at the recruiting station level within the areas of testing, counseling, and assignment. These techniques will ameliorate some of the problems recruiters and Navy manpower managers presently have in those area. This report has briefly described present efforts within project CONTRACT as regards the selection of a vehicle to implement those techniques. The vehicle selected is a functionally distributed computer network system. That system is characterized by the location of minicomputers at each of the Navy Recruiting Area and District offices and a central multi-processor located at Great Lakes. Unique to the system is the concept of mobile counseling vans, complete with a processor and CRT terminals. The mobile units should permit a greater access to the recruit applicant pool. Overall the system should permit a more thorough occupational selection process by the recruit applicant.

Of fundamental importance to project CONTRACT is the development of a prototype computer network that will be utilized by NPRDC to serve as the interface for implementing research techniques. This will result in a direct line for the implementation of techniques for utilization with that environment. The project consumer will now have a means for assessing the merits of implementing a particular technique without having to resort to the cost of a large scale implementation effort for a system of unproven value in the operational world.

REFERENCES

Hawkins, M. D., Crow, R., & Haltman, H. Estimates and implications of alternative configurations of an advanced personnel procurement system (APPS) (AFHRL TR 74-82). Lackland Air Force Base: Air Force Human Resources Laboratory, 1974.

Horst, P. & Sorenson, R. C. Matrix transformations for optimal personnel assignments. Navy Personnel Research and Development Center, San Diego. In progress.

Impelliterri, J. T. A computer-assisted career exploration system. National Business Educational Quarterly, 37, 1968, pp 47-51.

Katz, M., Chapman, W., & Godwin, W. SIGI-A computer-based aid to career decision-making. EDUCOM Bulletin, Summer Issue, 1972, 6p.

Rafacz, B. A. A survey of the developmental research for project CONTRACT. A presentation made to the 17th Military Testing Association Conference, Indianapolis, Indiana, September 1975.

Rafacz, B. A. & Halstead, D. The sequential assignment problem: phase I - the development of a computer simulation program to evaluate a sequential assignment strategy. Navy Personnel Research and Development Center, San Diego. In progress.

Underwood, M. & Rafacz, B. A. A feasibility study of a distributed network for navy recruiting. Navy Personnel Research and Development Center, San Diego. In progress.

Ward, J. H. Use of a decision index in assigning Air Force personnel. Lackland Air Force Base, Texas: Air Research and Development Command, April, 1959. (ASTIA Document No. AD 214 600)

Yellen, T. M. I. Computer-assisted occupational counseling: a prototype model. A paper presented at the 18th annual conference of the Military Testing Association, October, 1976.

Yellen, T. M. I. A computerized guidance and occupational system for civilian and related navy occupations. Navy Personnel Research and Development Center, San Diego. In progress.

DATA PROCESSING PROBLEMS IN COLLECTING JOB SURVEY INFORMATION

Johnny J. Weissmuller,

Air Force Human Resources Laboratory

Britt Kauffman,

University of Texas, Austin

DATA PROCESSING PROBLEMS IN COLLECTING JOB SURVEY INFORMATION

INTRODUCTION

Many data processing problems may be avoided by the improved design of data collection instruments. Improved design, however, necessitates a thorough understanding of both the input requirements and the operating characteristics of the processing systems to be used. Normally it is not feasible for the person preparing the instrument to be totally familiar with the computer programming packages, and it is therefore essential that the instrument designer locate and communicate with the people who do have this knowledge. Depending on the diversity of processing required, this may entail identifying and contacting several distinct groups of people in the data processing department. When this communication fails to take place, design decisions are made on criteria which are thought to be arbitrary, but which actually have significant processing implications. Moreover, the resulting problems are not even anticipated by the data processing department who becomes aware of the problems only after the operational processing has begun. By this time, problems which would have been easy to correct or prepare for have become major obstacles in extracting the desired information. Major obstacles, though usually not insurmountable, do exact a high price not only in computer time, but also in stress-filled man-hours spent by the programmer and those frustrating delays experienced by the waiting analyst.

GENERAL DESIGN CONSIDERATIONS

On the highest level, the prime considerations in the design of data collection instruments should be both economy and reliability. Normally in data processing clear trade-offs exist between these goals, but through technological advances coupled with good design, both of these objectives can be maximized without sacrifice. For example, in the area of occupational analysis many organizations are currently in the process of converting their initial data collection/entry from keypunched inventories to optically scanned response sheets. It is hoped that this change will reduce the total overall cost while not impairing the reliability. To insure that this hope can be realized, it is imperative that the criteria for the design of opscan sheets be considered at this time. A good design is only a good design with respect to a given context, and in this paper we deal only with the design of opscan sheets for use in occupational analysis.

Although we are restricting our context in order to provide specific examples, the principles of the analysis are general. To collect the information required for an adequate design, the following steps are required:

1. Define the context of the design,

2. Identify the systems and sub-systems involved,

3. Analyze the input requirements of each system, and

4. Seek information on both the ideal and actual operating characteristics of each.

SPECIFIC DESIGN CONSIDERATIONS

At this level in defining our context it is appropriate to identify the unique characteristics of job survey information as opposed to a knowledge test or an opinion survey. Historically, job survey instruments have evolved into task check lists in which job incumbents indicate only those tasks which they actually perform. In order to identify similarities across jobs, these check lists, or task inventories, grew in scope until an incumbent would respond to perhaps one out of every eight tasks in a typical survey. This methodology decreased the cost and increased the speed of collecting and processing job information.

The significant point is, however, that the job incumbent is selecting a relatively small number of items from the full inventory. This characteristic, low response density, creates unique problems and opportunities for both the processing and collection of data. For processing this type of data, computer programs can be written which capitalize on this sparse response characteristic. In fact, the package of computer programs with which we will deal is designed based on this assumption. We will be referring to the Comprehensive Occupational Data Analysis Programs (CODAP) system as developed by the U.S. Air Force Human Resources Laboratory and implemented by several government agencies and universities. Because the CODAP design is based on this assumption, it is very efficient for processing job survey information, but conversely it is rather cumbersome and costly to process a survey in which each respondent is required to answer every question. In the area of data collection this low response density can be either a time-saving advantage or a frustrating disadvantage, and the design of the data collection instrument itself makes the difference. This point will be elaborated in the discussion of various implementations and the Case History which follows it.

SYSTEMS AND SUB-SYSTEMS IN JOB SURVEY INFORMATION HANDLING

In the most general sense, there are four major systems involved in handling job survey information. These systems may roughly be defined as follows:

1. Instrument Construction

2. Survey Administration

3. Data Processing

4. Survey Analysis

Figure 1 demonstrates how each of these systems contains people who are responsible for the assumptions and daily operations. Key people in each of these systems can usually provide valuable information to aid in the design of the survey instruments. As data processing specialists, we, the authors, have been approached by various people within the data processing department regarding job survey instruments. Because of these initial contacts we decided to actively solicit more comments to be included here as a sample of data processing suggestions for increased efficiency.

In soliciting information about specific data processing problems, a survey designer may make an honest attempt at communication, but find himself talking with a totally disinterested (or confused) party. This is the result of the fact that non-data processing personnel usually fail to realize that there are distinct areas of specialization within data processing. At the very minimum there are four specialities involved in the processing occupational data. The first of these areas is that of designing the programs and procedures for use with the optical scanning equipment. Because this is a relatively new frontier, the person in this area will usually be a very sophisticated programmer/analyst who is used to accepting new challenges. The second area is that of actually operating the optical scanner on a daily basis. In this area, one finds people who are apt to know all the things which can retard processing; things totally unsuspected by the above programmer or the optical scan machine serviceman. The third specialization encompasses the programmer/analyst responsible for the design of the computer analysis programs. In our context, he would be the programmer who maintains and upgrades the CODAP system. The fourth area is that of operational studies. The people in this area routinely accept tapes or punch cards produced by the optical scan department, correct the detectable errors, and provide the survey analyst with the requested reports using the CODAP system. These CODAP applications programmers are usually familiar with conversion and utility programs outside of the CODAP system because the optical scanner tapes require pre-processing prior to entering the data into CODAP. Depending on the size of the organization, a single person may act as one or more of these four specialists. The larger the organization, the more likely that each area will have distinct people or groups. We contacted this range of data processing personnel to solicit observations or suggestions and the following sections contain their responses.

A. OPTICAL SCANNING: PROGRAMMER/ANALYSTS

From the programmer/analysts who design the optical scanner programs and conversion procedures we received a request to avoid mixing alphabetic and numeric characters together in the same response columns. If that is not possible, at the very least, point out this condition before the translation programs are written. Secondly, avoid splitting a single character response into multiple columns or rows. If, for example, a single digit response of 1 to 9 is solicited on three rows (1, 2, 3), (4, 5, 6), and (7, 8, 9) special programming is required to make the conversion program understand that the three rows represent a single response.

B. OPTICAL SCANNING: DAILY OPERATORS

The operators of the optical scan equipment had several suggestions. First, if specially designed and over-printed optical scan sheets are being used, test batches of these forms should be run on the scanner before these forms are mass produced and sent out to the field. These test runs are invaluable for detecting such problems as missing or improper page codes, improper trimming of the booklets, and unnecessary "dark marks". An excessive "dark mark" in the center of a scan sheet can cause some scanning equipment to halt and require operator intervention. Such "dark mark" conditions have been caused by organizational emblems printed on the cover sheets of survey booklets. The service personnel who do maintenance on the scanner point out that such massive dark areas overload the sensors and cause the machine to halt, but they seem to be at a loss to explain why a fine-lined "X" across an entire page would have the same effect. Well-intentioned survey designers have been known to have such "X" marks printed on pages along with instructions to incumbents to not mark on that page. Thus far we have been lucky that the reverse side of these offending pages has not contained data because the scanner attempts to read both sides of a sheet simultaneously, and the operators, to avoid the halts, remove these sheets prior to the actual scanning.

Though the following observations are not the result of the survey design itself, the optical scan machine operators believe that additional instructions on the booklets may help to reduce these problems. The second item which can delay processing is that of careless handling of the scan sheets. When sheets with bent or torn edges jam in the optical scanner, they are usually mutilated requiring the operator to transcribe the response to a new sheet. A related third point is that the "optical scanner" actually scans in the infrared part of the spectrum. Number two pencils make marks which are detected very well, but even the darkest of marks made with felt tip pens are totally invisible to the scanner. Again, when this type of violation occurs and is detected, the optical

scan operator must remark all responses in the entire booklet using a number 2 pencil. Finally, stray marks on the forms can also cause the machine to halt. The most sensitive areas of the form are usually the top and the timing edges. Oddly enough, job incumbents appear to be rather co-operative in this area, but, for some reason, the survey proctor or administrators do not seem to believe that this prohibition applies to them. Many times such comments as "Batch 1" or "These are all the booklets" have to be erased or snow-painted before the machine will process them.

C. CODAP: PROGRAMMER/ANALYSTS

The programmer/analysts who are responsible for the CODAP system have two observations which may be of interest. These observations may be summarized as the following cautions:

1. Avoid ambiguous coding conventions, and

2. Avoid low response density when possible.

Under certain conditions, a computer program may interpret both blanks and zeroes as zeroes which can be disasterous if separate meanings are intended. This problem has occurred in soliciting a YES/NO response in which a ZERO represented "yes" and a non-response (BLANK) meant "no". This YES/NO choice is usually employed to identify "equipment used" or "courses taken" from a large list. Although using ONE for "yes" and ZERO for "no" would be ideal since it also avoids the low response density, to reduce the marking required for very long lists, it is acceptable to use a non-response (BLANK) for "no". In this way the ambiguity may be used to an advantage. By computing the mean value of ONE and BLANK (zero), the resulting value may be immediately interpreted as "percent of the group" responding yes by simply moving the decimal point two places to the right.

As an example of how both these cautions can be violated in one design step, consider the use of multiple variables to determine a single attribute. This probelm actually occurred in a survey where a 66-valued attribute (coded job title) was solicited on a standard 80 column opscan sheet. Since the scan sheet had 80 columns, each of which had possible responses of BLANK and 0 to 9, the decision was made to solicit the coded job title as seven separate responses. For each of these seven questions, job titles were associated with each of the values 0 through 9, and a non-response (BLANK) indicated "none of the above" for this category. Because of this decision, severe processing problems have been created for the CODAP applications programmers. For each of the seven categories, the BLANK and ZERO responses are indistinguishable for certain programs. Consequently,

the "ZERO" responses for each category includes not only the people who actually responded with a ZERO, but also every person who left that column blank while marking one of the other six. Moreover, because the incumbent was responding to only one out of seven, identically coded columns (0-9), he had the additional problem of locating the proper column for his response. If he inadvertently responded in the wrong column, it is not likely that he caught the mistake because he expected "some" blank columns on either side of his response.

In this case the response density could have been increased by collapsing the seven questions into a single question with a two digit response. By coding the job titles from 01 through 66, the incumbent would have to fill in both columns, maintaining a high response density, and at the same time eliminating the confusion caused by having seven categories to determine a single attribute. From a processing standpoint it is much easier to identify and select subsamples when only one variable needs to be considered.

It is important to note that this category confusion by the incumbent can be caused without actually separating the attribute into distinct questions. In another case, a three column display of supposedly mutually exclusive choices was given in the questionnaire with a corresponding three column response grid on the opscan answer sheet. Some incumbents interpreted this configuration to imply that multiple responses were acceptable and, by stretching their imaginations a little, qualified themselves to fill in the bubbles in two or three columns. Unfortunately, the opscan conversion program was told that the three columns represented only a single response and every multiple response was marked invalid. Although the survey analyst was able to formulate a decision logic for resolving the multiple responses, the optical scan conversion program had already collapsed the multiple responses down to a single asterisk, regardless of which codes had caused the conflict. By the time the tape reached the CODAP applications programmers, the data needed to apply that decision logic had already been lost, and the deadlines precluded any re-scanning efforts.

D. CODAP: APPLICATION PROGRAMMERS

The application programmers who process CODAP studies asked us to reiterate the need for improved communications. The problems mentioned above could have been avoided or minimized had communication been better, but the following example is one in which the lack of communication caused the problem. This classic example comes from a survey in which "length of time on the job" was being solicited in months as a three digit response, but the printed answer sheet only allowed two digits. This error was detected after the forms had been printed but before the inventory was administered. The survey designer

simply instructed the administrators to have the job incumbents record their responses in three previously unused columns at the end of the answer sheet. This would seem to have solved the problem. The decision to change the columns on the answer sheet, however, was not relayed to the person responsible for preparing the data for input to the CODAP system. In the absence of directions to do otherwise, he used the two original columns specified on the form. Later, having noticed the discrepancy, he assumed that the data had effectively been lost due to the fact that people with more than 99 months on the job could not record their response. In talking with the project director one day, he mentioned the loss of the data, and only then was he informed of the additional instructions. By this time several hundred dollars had been spent in processing this data and it was, therefore, prohibitive in cost to simply start over to recover the missing information. The data was eventually recovered after special processing and several days of delay. Luckily, in this case, the opscan system used punched cards for all columns on the answer sheet. There are systems which only record data for the specified areas. Had one of those systems been used in the last example, all the opscan sheets would have had to been reprocessed, either through the optical scanner or manually by keypunch personnel.

As part of this improved communication, the CODAP applications programmers would like to have the opportunity for a post-printing, pre-administration review of intended surveys. In conjunction with the optical scan operators test batches, the applications personnel can help avoid unintentional errors in content. Many times they will encounter one or two duplicated task statements or a task statement which is missing from an inventory which is supposed to exactly match another. These errors can be detected by the computer in the process of automating the titles, and, if this is done prior to the administration of the survey, the incumbents can be alerted to the problem and instructed to take appropriate actions.

IMPLEMENTATION OPTIONS

With all of these points in mind, let us now review several implementation strategies. Having observed several implementations of the optical scan technology, we have identified three general approaches in the occupational analysis area. The first, and least expensive type, employs a standard "off the shelf" 80 column opscan response sheet and a separate questionnaire booklet. The questionnaire booklet indicates the page and column number in which to record both control and response information. The second type also employs a separate questionnaire, but uses a specially designed opscan response booklet in which the number of the response column is the same as the number of the task in the inventory. Background and control information are usually recorded in an area distinct from the normal

task responses. The third approach is to use a single booklet in which the task titles are overprinted directly on the opscan form, adjacent to the appropriate response column. Again, as in type two, as much control information as possible (pages, codes, etc.) is pre-recorded in the over-printing process, but whatever control or background information the incumbent is responsible for, is recorded in an area separated and distinct from the task responses.

A. COSTS

The first approach entails the least direct cost. A standard form is used which is relatively inexpensive and easy to obtain. The printing of the corresponding questionnaire booklet requires no outstanding technical considerations. The second approach also uses the straightforward questionnaire booklet, but is uses specially designed opscan response booklets rather than standard forms. these specially designed opscan forms incur additional expense not only for the initial design costs, but also for the recurring costs of producing forms according to stringent technical requirements. These opscan forms are printed almost exclusively by manufacturers of optical scanning equipment and the major forms producing companies. In the third approach, the separate questionnaire booklet is dropped, but instead the blank opscan forms are sent to a local printer who overprints the control information, instructions, and task statements, then cuts and binds the sheets to form a single document. The local printer who is responsible for printing the task titles and page codes on the opscan forms must be aware of the stringent requirements on these forms. Not only must the critical spacing between page codes be maintained, but after the sheets are printed, care must be taken to avoid cutting or trimming the control edges on these forms. A slight trim on the wrong edge of the sheet can render entire booklets unprocessable. If these sheets are bound together with staples rather than glue, the machines which sever the binding for entry into the optical scanner may experience many nicked or dulled blades. The only safe way to insure that these type three booklets are acceptable is to actually process a sample batch before the surveys are mass produced and sent out into the field.

Aside from these direct costs, there are indirect costs to be considered. Generally speaking, a more sophisticated questionnaire format will require less time and trouble for a job incumbent to complete. Hopefully, this increase in direct cost buys an increase in reliability and a reduction in man-hours required of the organization being surveyed. For example, when using an "off the shelf" 80 column opscan sheet, respondents are directly responsible for coding their booklet number and sheet number on each page. An error in this control information will not be detected by the optical scan processing, but it will cause the entire case to be dropped from any CODAP analysis. Particularly in CODAP studies where job-typing is to be done, a

systematic sampling error can be introduced if a certain group or groups of people tend to make errors in control information. In the Air Force version of CODAP there is a program which is devoted to detecting and identifying such errors so that the data may be corrected whenever possible.

B. OPPORTUNITY FOR TRANSCRIPTION ERRORS

On a type three booklet, once the incumbent selects his response, he merely records it in the adjacent column with no searching required. However, on types one and two, the incumbent, after deciding on his response, must search separate answer sheets for the proper location in which to record his response. For a type two booklet, this search is a simple matter of finding the column number which corresponds directly to the task number. We are assuming, of course, that the users of CODAP have stopped numbering the tasks from 1 to N within each duty, and begun using the new system of numbering the tasks from 1 through the number of tasks in the full inventory, regardless of duty. To respond to a type one booklet, the incumbent is required to locate a particular coding sheet and a particular column whose number may never be the same as the task number.

C. A CASE HISTORY

A "pure" type one approach is rarely used. We have, for example, seen a modified type one approach in which the control information was pre-coded by clerical personnel, and the coded sheets were bound together to form a booklet. In addition, for the few coding forms on which background information was solicited, titles were printed over the columns reserved for those items. With the exception of the fact that the forms still had the standard 1 to 80 numbering on each sheet, this implementation approximated a type two approach. This difference, however, is still significant, particularly when the control information is coded in the first five columns, leaving columns numbered 6 to 77 available for responses.

In an attempt to develop a quantitative measure of the transcription or "parallax" problem, we designed a series of programs to evaluate four inventories. These four, modified type one surveys solicited both relative time spent and a secondary factor. These responses were to be coded on separate sets of answer sheets and each set was to be completed before starting the next. By the definition of a secondary factor, if an incumbent indicated a relative time spent for any task, he was required to respond to the secondary factor for that task also. Because of this, we were provided with a machine sensible way to detect discrepancies. We processed these inventories, listing out all tasks which were missing one of the two responses. In this initial step no attempt was made to distinguish errors of omission from the errors of transcription. The data from this

first step are recorded in Figure 2. This data implies that about one out of every four respondents are making errors and that, for these respondents, the errors account for 5 to 8% of all the information that they are reporting. It is difficult to say whether these figures represent the upper or lower limits of the problem. To the extent that the errors are caused by isolated transcription errors, the figures may be slight over-estimates. Since an error is defined as a task with one rating missing, a person who meant to respond to only two tasks, but misplaced the secondary rating, would show an error rate of 66%, rather than 50% as would be expected. Conversely, to the extent that the errors are caused by blocks of transcription errors, the figures may be under-estimates. In one instance a block of 36 consecutive responses were found in which all the secondary responses were shifted down one column. Is the error rate here actually 5.5% as the algorithm indicates, or is it really 100%? Note also that only discrepancies were detected. Any errors which were systematic went undetected, but systematic errors should be uncommon, correct?

While we were attempting to distinguish between missing responses and shifted responses, several cases were dropped from consideration because of their extremely high error rate and the apparent senselessness of the errors. After all the other cases had been processed, we returned to investigate these cases. Most of them turned out to have a shift error of not one, but six columns. We could not explain why there would be a six column shift among the responses, but we did note that this six column error was almost exclusively restricted to the first opscan coding sheet. At first we surmised that the respondents had simply failed to record the control information in the first five columns and recorded their task responses beginning in column one. This would only account for a shift of five, however. This hypothesis was partially untenable due to the fact that CODAP would drop any case which had an error in the control information, and all the cases studied had already been processed in CODAP. Upon considering the input to CODAP, we noticed that the control information only allowed three digits for case identification and CODAP required four. We then found that some pre-processing had been performed on the data to expand the case control number to four digits. This meant that the control information could account for the entire six-column shift. It was at this point that we learned that the control information had been pre-coded and the incumbent never needed to concern himself with the first five columns of his answer sheet. For someone who did not have to respond to the first five tasks, it seemed only natural to mark his response in the corresponding column number. These "natural instructions" fail when the second coding form is reached which is also numbered 1 to 80. When all else failed, he read and followed the actual instructions. Because of the low response density a respondent expects "some" number of blank columns around his answers and he loses the visual keys or feedback that would help him to detect and correct his errors.

In our analysis of this problem, less than 1% of the people in three of the four surveys had this six-column shift. In the fourth survey, however, this error accounted for over 12% of the population. It should be stated again, these were only the detectable discrepancies. How many people made this error on the first set, realized the mistake while coding the second set and at that point corrected set one? How many people followed their "natural instructions" for set two also, and went undetected? For the people who discovered the error because of the second set, what if there had not been a set two? There are perhaps more questions here than there are answers, but if one values reliability, these are the questions which must be answered.

Perhaps the above problem could be solved by coding the control information in the last five columns of each answer sheet. That way not only could the incumbents use their "natural instructions" for the first 70 or so tasks, if they attempted to respond to task 75, they would find the pre-coded control information there and possibly decide to read the directions.

CLOSING

We hope that we have provided information that may be of use in the design decisions for job survey instruments. We've tried to address those areas in which improved design or communication could have a favorable impact, but there will always be some areas over which we have no control. We, of data processing, have felt that we had no control over the design of the opscan forms, but even lacking the direct control, communication can help solve the problems. We are the ones who experience the consequences of your design decisions, and we feel it is our responsibility to provide feedback in order that your future designs may build upon this knowledge.

JOB SURVEY INFORMATION SYSTEMS

I. SURVEY INSTRUMENT CONSTRUCTION

- A. CONTENT SPECIALIST
- B. DATA COLLECTION INSTRUMENT SPECIALIST
- C. OPTICAL SCAN SHEET MANUFACTURER
- D. LOCAL PRINTER

II. SURVEY ADMINISTRATION

- A. PROCTOR OR ADMINISTRATOR
- B. JOB INCUMBENT

III. SURVEY DATA PROCESSING

- A. PROGRAMMER/ANALYST FOR OPTICAL SCAN MACHINE
- B. OPERATOR OF THE OPTICAL SCAN MACHINE
- C. PROGRAMMER/ANALYST FOR COMPUTER ANALYSIS PROGRAMS
- D. APPLICATIONS PROGRAMMER

IV. SURVEY ANALYSIS

- A. SURVEY ANALYST
- B. RESEARCH ANALYST

Figure 1

ERROR ANALYSIS OF FOUR, MODIFIED TYPE ONE, JOB SURVEYS

	Number of Cases in the Total Sample	Number of Cases with any error	Percent of All cases which contain any errors	Number of Tasks in the Inventory	Average Number of Tasks Performed	Only Cases with Errors - Average Percent of Tasks in error	Total Sample - Average Percent of Tasks in error
SURVEY A	357	36	368	94	25.5	7.81	1.99
SURVEY B	675	56	596	130	21.8	4.99	1.09
SURVEY C	418	62	860	282	32.7	4.83	1.58
SURVEY D	450	91	114	96	84.2	7.87	6.63

Figure 2

RELATIONSHIPS BETWEEN ATTITUDES TOWARD COMPUTER-ASSISTED INSTRUCTION AND TRAINING PERFORMANCE

Bruce W. Knerr

U.S. Army Research Institute for the Behavioral and Social Sciences

Student attitudes toward training media, particularly, Programmed Instruction (PI), Computer-Assisted Instruction (CAI), and Computer-Managed Instruction (CMI), have long been a topic of interest for both Army trainers and researchers. Army trainers typically assess student attitudes toward CAI or CMI systems as a part of the evaluations of those systems. The evaluation of the PLATO IV CAI system at the Army Ordnance Center and School and the proposed evaluation of the Computerized Training Systems Project ABACUS at the Army Signal School are recent examples of this. Interest on the part of researchers is evidenced by the substantial number of studies which have collected student attitude data. King, in a 1975 review of the research literature in this area, cited 44 such studies.

However, this interest has produced neither well-integrated nor consistant results, particularly with regard to relationships between student attitudes and student performance in the instructional situation. King (1975) noted two methodological problems which, it would appear, are at least partially responsible for this lack of integration and consistency. First, while student attitude data are frequently collected (as noted above), there has been no effort to investigate student attitudes in any integrated fashion. Only rarely are student attitudes the primary focus of the research. Investigators, justifiably, are more concerned with the effects of the independent variables on cognitive or psychomotor performance than on affective measures. Second, there is little agreement on an operational definition of student attitude. There have been some attempts to achieve consistency in the use of measuring instruments: five of the studies cited by King, and one additional study to be cited here (Gallagher, 1970), have used either a scale developed by Brown (1966) or modifications thereof. For the most part, however, the scales

---

The findings in this report are not to be construed as official Department of the Army position, unless so designated by other official documents.

used are designed ad hoc with neither the items used or the metric properties of the scales described. As King noted, "most studies use experimenter-constructed tests which have unknown or unreported reliabilities (1975, p.7)."

Relatively few studies have attempted to investigate what, if any, relationships exist between attitudes toward PI, CAI, or CMI and performance in a course of instruction taught by one of those methods. Those that have, have produced mixed results. Doty and Doty (1964) found posttest attitude (that is, attitude measured following the instruction) toward PI to be positively correlated with lesson achievement for female students, but not for male students. Gallagher (1970) found no significant relationships between posttest attitude toward CMI and course performance. Mathis, Smith, and Hansen (1970) measured attitudes toward CAI both before and after the instruction was presented. Errors per question attempted during the course of the instruction correlated negatively with posttest attitude, but not with pretest attitude. Reid, Palmer, Whitlock, and Jones (1973) measured attitudes toward CAI before and after the instruction was presented. Pretest attitude, posttest attitude, and attitude change were unrelated to lesson achievement and to the amount of time required to complete the lesson for the total sample. While significant correlations between attitudes and performance were obtained for selected subgroups, it appears that no more than 5.2% of the correlations calculated were significant at the .05 level, and thus could be attributed to chance alone.

The above studies were conducted using college or graduate students, and it was felt that the attitude scales used were not appropriate for use with enlisted military personnel. Consequently, Knerr and Nawrocki (1976) developed two attitude-toward-CAI scales, a 13-item pretest and a 37-item posttest, for use with military personnel. When used in the evaluation of the PLATO IV CAI system at the Army Ordnance Center and School, neither the pretest nor the posttest was correlated with the time required to complete a series of four lessons presented by CAI, the only performance criterion available.

The purpose of this study was to continue the investigation of the relationships between military student attitudes toward CAI, and student performance begun by Knerr and Nawrocki (1976) using a variety of measures of performance in the instructional situation.

Method

Subjects

Subjects were 90 enlisted military personnel receiving training at the U.S. Army Engineer School (USAES), during the period September - November 1975. A description of salient student background information is provided in Table 1. Median values show that the "typical" student was male, in pay grade E-2, 19 years old, had been in the military for 3 months, had 12 years of formal education (all were at least high school

872-A

TABLE 1

STUDENT BACKGROUND INFORMATION

	Median	Range
Pay Grade	E-2	E1 - E5
Age	19	17 - 32
Months in Service	3	2 - 53
Years Education	12	12 - 16
Typing Speed (words per minute)	0	0 - 100

	Male	Female
Sex	81	7

Note. Background data on two students was destroyed by systems failure. N = 88.

graduates) and could not type.

Attitude Scales

The two attitude scales, pretest and posttest, developed by Knerr and Nawrocki (1976) were used. The pretest consists of 13 five-alternative multiple choice items. When previously administered, it was found to have an internal consistency (Kuder-Richardson Formula 8; KR 8) of .91. The posttest consists of 37 items, and has an internal consistency of .89. It can be divided into two subtests. The first, the repeated subtest, consists of those 13 items in the pretest, but rephrased in the past tense where necessary. The repeated subtest has an internal consistency of .85. The unique subtest consists of 24 additional items, and has an internal consistency of .81. Sample items are shown in Figure 1.

The scales had previously been administered in paper and pencil format. Prior to this study they were modified for on-line administration and scoring by the PLATO IV CAI system.[1] The experimental procedure was such that the scales had to be administered on an individual basis. Computerized administration, then, offered a substantial time savings for the experimenter. It also provided the capability for automatic scoring and data reduction.

Figure 2 shows a sample item as it appeared initially to the student. The student responded by entering the letter of his or her choice (a through e) via the terminal keyboard. This choice was then marked by an asterisk, and the student was given an opportunity to change the answer before advancing to the next question, as shown in Figure 3. The student was required to answer each question in order to advance to the next.

Lesson Material

The lesson material consisted of a four-section lesson entitled "The Elementary Physiology of Audition (or How Your Ear Works)." The sections dealt with sound waves, anatomy of the ear, balance, and ear damage and degeneration, respectively. Following each section, the students received a series of review questions covering the objectives of that section. Students who answered less than 85% of the review questions for any section correctly were required to repeat that section. After the fourth review was successfully completed, the student was administered a 68-item test covering the objectives of the entire lesson.

Performance Criteria

The following performance criteria were obtained for each student: (a) the amount of time required to complete the four lesson and review

[1]Commercial designations are used only for precision of description. Their use does not constitute endorsement by the Army or the Army Research Institute.

Pretest

I would prefer to have most courses taught by computer rather than by other teaching methods.

Taking a lesson taught by computer would be too mechanical.

Repeated Posttest

I would prefer to have most courses taught by computer rather than by other teaching methods.

The lessons taught by computer were too mechanical.

Unique Posttest

Based on my experience with this lesson, I prefer CAI to other methods of instruction

The method by which I was told whether I had given a right or wrong answer became boring.

Response alternatives for each of the above items are:

a. Strongly agree
b. Agree
c. Undecided
d. Disagree
e. Strongly disagree

Figure 1. Sample scale items.

Most courses could be taught more effectively by a regular teacher than by computer.

a. Strongly agree
b. Agree
c. Undecided
d. Disagree
e. Strongly disagree

802

Figure 2. Pretest item 1, as it appears to the student before he selects his response.

Most courses could be taught more effectively by a regular teacher than by computer.

a. Strongly agree
* b. Agree
c. Undecided
d. Disagree
e. Strongly disagree

803

b

Your answer is marked with an '*'. You can change it by pressing -BACK-, or continue with the next question by pressing -NEXT-.

Figure 3. Pretest item 1, as it appears to the student after he has selected his response.

sections (time); (b) the total score on the final end-of-lesson test (total score); (c) the total number of incorrect responses made during the lesson and review sections (errors); (d) the percentage of the total responses made during the lesson and review sections that were incorrect (percent errors); and (e) the number of reviews failed (review failures). Percent errors requires some additional explanation. It is the percentage of responses to which the student did not receive positive feedback (OK, correct, right, etc.). It is also proportional to errors per question attempted, which Mathis et al. (1970) found to be correlated with posttest attitude.

Experimental Procedure

Upon arrival at the experimental area, each student was given a short written explanation of the experimental procedures. After any questions were answered he or she was signed on to the PLATO IV system. The PLATO IV system controlled all student routing, lesson administration, and data collection. Unless a problem arose, the student did not interact with the experimenter again until the end of the session.

After the student was signed onto the system, he or she received a short lesson on the use of the student keyboard and terminal. This was followed by an on-line biographic questionnaire, the attitude pretest, the human audition lesson, the end-of-lesson test, and the attitude posttest.

Results

The results have been organized around three central issues. The first is the internal consistencies of the attitude scales. The second is the relationships between the performance criteria and the attitude scale scores. The final issue is the relationships between attitude change and the performance criteria.

Internal Consistency

Table 2 presents a comparison of the internal consistencies of (KR 8) of the on-line versions of the pretest, total posttest, and unique and repeated subtests with the internal consistencies of the previously administered off-line (paper and pencil) versions. The internal consistencies for the on-line versions were .88 for the pretest, .93 for the repeated posttest, .87 for the unique posttest, and .93 for the total posttest. Each of these values was compared with the internal consistency of the same scale when administered in paper and pencil format. Differences were tested using the z test for the significance of the difference between two correlations. The results of these tests are also shown in Table 2. While it initially appeared that the on-line versions of the posttest had higher internal consistencies than did the paper and pencil versions, this was supported only for the repeated posttest, for which the on-line version had an internal consistency of .93, and the off-line version, .85 ($z = -2.40$, $p < .01$).

TABLE 2

A COMPARISON OF THE INTERNAL CONSISTENCIES (KR 8)

OF THE ON-LINE AND OFF-LINE VERSIONS

	Off-Line	On-Line[a]	$z_{diff.}$
Pretest	.91[b]	.88	1.91
Repeated Posttest	.85[c]	.93	-2.40*
Unique Posttest	.81[c]	.87	-1.20
Total Posttest	.89[c]	.93	-1.44

[a] $N = 90$

[b] $N = 228$

[c] $N = 64$

*$p < .01$

Attitude-performance Relationships

The matrix of intercorrelations among the attitude scale scores and the lesson performance criteria is shown in Table 3. The left side of the table indicates that the performance criteria are highly correlated, with the absolute value of the correlations ranging from .57 to .90. High values of total score reflect good performance, whereas high values of the other performance criteria reflect poor performance. Thus the correlations between total score and the other performance criteria are negative.

The lower left corner of the intercorrelation matrix shows the intercorrelations among the various attitude scales. Correlations of the pretest with the repeated, unique, and total posttests are .78, .59, and .71, respectively. The correlation between the unique and repeated posttests is .78.

The attitude pretest appears to have no predictive validity. Correlations of the pretest with the performance criteria range from -.02 to .09. None of these correlations is significant at the .05 level.

Two of the performance criteria showed significant relationships with the post-lesson attitude scales. Total score was correlated with the total posttest ($r = .18$, $p < .05$) and the unique posttest ($r = .21$, $p < .05$). Percent errors was correlated with the total posttest ($r = -.22$, $p < .05$), the unique posttest ($r = -.22$, $p < .05$), and the repeated posttest ($r = -.19$, $p < .05$).

Attitude Change and Performance

The relationships between attitude change and the performance criteria were also determined. The measure of attitude change used was residual change, that is, that portion of the posttest score which is independent of the pretest score. Use of residual change in this situation has two advantages over the use of a raw gain score (posttest minus pretest): (a) it is not correlated with pretest score, while raw gain is negatively correlated with pretest score; and (b) it does not require that the prestest and posttest scores have the same metric (Manning & DuBois, 1962). Correlations between the performance criteria and residual change were calculated by using the part correlation (McNemar, 1962, p.167). Thus the correlation calculated is that between a performance criterion and a posttest score with the effects of the pretest "removed" from the posttest score. Three different residual changes were calculated: pretest-unique posttest, pretest-repeated posttest, and pretest-total posttest.

The resulting correlations are presented in Table 4. Time and review failures were not correlated significantly with any of the residuals. The remaining three criteria (total score, errors, and percent errors) were correlated with each of the residuals. The absolute value of these significant correlations ranged from .18 (errors and the

TABLE 3

CORRELATIONS BETWEEN THE ATTITUDE SCALES AND THE LESSON PERFORMANCE CRITERIA

	Total Score	Errors	% Errors	Time	Review Failures	Pretest	Repeated Posttest	Unique Posttest	Total Posttest
Total Score	1.00	-.62**	-.59**	-.64**	-.63**	-.02	.12	.21*	.18*
Errors		1.00	.90**	.65**	.79**	-.00	-.12	-.17	-.15
% Errors			1.00	.57**	.66**	-.06	-.19*	-.22*	-.22*
Time				1.00	.76**	-.02	-.06	-.12	-.10
Review Failures					1.00	.09	-.02	-.13	-.08
Pretest						1.00	.78**	.59**	.71**
Repeated Posttest							1.00	.78**	.94**
Unique Posttest								1.00	.95**
Total Posttest									1.00
Maximum N	90	87	87	90	89	90	90	90	90
Mean	79.31	51.84	20.72	106.79	1.53	45.91	47.42	91.86	139.28
Standard Deviation	9.44	51.19	9.65	34.57	1.74	8.34	10.25	11.86	20.86

*p < .05
**p < .01

807

883

884

TABLE 4

CORRELATIONS BETWEEN RESIDUAL ATTITUDE CHANGE AND LESSON PERFORMANCE CRITERIA

Performance Criterion	N	Type of Residual		
		Pre-Total	Pre-Repeated	Pre-Unique
Total Score	90	.28**	.22*	.28**
Errors	87	-.19*	-.18*	-.20*
% Errors	87	-.25*	-.22*	-.23*
Time	90	-.12	-.08	-.13
Review Failures	89	-.03	.07	-.09

808

*p < .05
**p < .01

pre-repeated residual) to .28 (total score and the pre-total and pre-unique residuals). Thus the proportion of variance in attitude change accounted for by the performance criteria varied between 3.24% and 7.84%.

Discussion

This study has demonstrated the feasibility of using on-line scales to measure military student attitudes toward CAI. Certainly there are many questions to be answered about such an approach. For example, it is not known whether on-line (as opposed to paper and pencil) administration has any effect on student responses. While there was one significant difference between the internal consistencies of the on-line versions and those of Knerr and Nawrocki's (1976) paper and pencil versions, the background characteristics of the students in the two studies were not exactly equivalent and, more importantly, the amount of exposure that the students had to CAI, both before the administration of the pretest and between the administration of the pretest and the posttest, differed for the two groups. Thus the difference obtained may be a result of factors other than the method of administration.

While acknowledging that these questions are still unresolved, on-line adminstration still appears to have certain advantages. The first is the reduction of personnel time required for administration and scoring. A second advantage is that the scales can be integrated into the computer system's management function, thus insuring that each student is administered the proper scale at the proper time. Finally, the method of presentation eliminates certain types of student errors, such as answering a question in the wrong space on the answer sheet, or failing to answer a question.

The internal consistencies of .88 for the pretest and .93 for the total posttest compare favorably with those of previously developed scales. Few other scales which could be used as pretests have reported reliabilities. Gallagher (1970) reported an internal consistency of .937 (Hoyt method) for the Michigan State University Learning Service "Attitude Toward Learning Inventory." While this scale is suitable for use as a pretest, it is not designed specifically to measure attitudes toward CAI. Mathis et al. (1970) developed a 30-item pretest, but reported no internal consistency for it.

Internal consistencies (KR 20 or equivalent) for other scales used as posttests are .89 for Brown's (1966) 40-item scale, and .82 for the 27-item scale developed by Mathis et al. (1970). Many scales, such as those used by Doty and Doty (1964) and Reid et al. (1973) have no reported internal consistencies.

Evidence is mounting that pre-instructional attitude toward CAI does not effect subsequent performance in the instructional situation, at least when relatively naive students are involved. Knerr and Nawrocki (1976) found this to be the case using time as the criterion. Mathis, Smith, and Hansen (1970) obtained similar results using errors per

question attempted as the criterion. The present study found an absence of predictive relationships using a variety of lesson performance criteria.

The caveat regarding naive students may be an important one. Most studies have used students who had no prior experience with CAI. In the current study, the experience of the students prior to the administration of the pretest was limited to the time required to complete the terminal familiarization lesson and the biographic information questionnaire, less than 30 minutes. Thus CAI was still a novel and somewhat unfamiliar situation for the students. It is not known if similar results would be obtained when students have more extensive previous experience with CAI. For example, if, following the administration of the posttest, students had taken a second CAI lesson, what would have been the relationship between the attitude posttest and performance in the second lesson?

It was discovered that certain aspects of student performance in the instructional situation are related to posttest attitude (total score and percent errors) and to attitude change during the course of instruction (total score, errors, and percent errors). However, these effects were not large. None of the performance criteria accounted for more than 5% of the variance in posttest attitude, or more than 8% of the variance in attitude change.

Implications

In view of the lack of predictive relationships between the attitude pretest and subsequent performance, measuring pretest attitude for naive students would seem to be of little value unless one is interested in measuring attitude change, per se. It would not, for example, appear to be useful for deciding whether a particular student should receive instruction by CAI or by some other medium. This may not be the case for students who have had extensive previous experience with CAI, however.

Preparation of CAI lesson materials which produce low student error rates during the course of instruction and on the post-lesson test may be benficial in terms of maintaining or producing positive student attitudes toward CAI. This must be considered tentative until such time as these factors are manipulated and the resulting effects on student attitude are measured.

References

Brown, B. R. An instrument for the measurement of expressed attitude toward computer-assisted instruction. In H. E. Mitzel and G. L. Brandon (Eds.), Experimentation with computer-assisted instruction in technical education. (Semi-Annual Progress Report, Project No. OEC-5-85-074), University Park, PA: The Pennsylvania State University, December 31, 1966).

Doty, B. A. and Doty, L. A. Programmed instruction effectiveness in relation to certain student characteristics. Journal of Educational Psychology, 1964, 55, 334-338.

Gallagher, P. D. An investigation of instructional treatments and learner characteristics in a computer-managed instruction course. (Tech. Rep. No. 12, CAI Center, Institute of Human Learning), Tallahassee, FL: Florida State University, 1970.

King, A. T. Impact of computer-based instruction on attitudes of students and instructors: A review (AFHRL-TR-75-4). Lowery Air Force Base, Colorado: Air Force Human Resources Laboratory, Technical Training Division, May, 1975.

Knerr, B. W. and Nawrocki, L. H. The measurement of military student attitudes toward computer-assisted instruction. ARI Technical Paper (in press), 1976.

Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Manning, W. H. and DuBois, P. H. Correlational methods in research on human learning. Perceptual and Motor Skills, 1962, 15, 287-321.

Mathis, A., Smith, T., and Hansen, D. College students' attitudes toward computer-assisted instruction. Journal of Educational Psychology, 1970, 61, 46-51.

McNemar, Q. Psychological statistics (3rd ed.). New York: Wiley, 1962.

Reid, J. B., Palmer, R. L., Whitlock, J., and Jones, J. Computer-assisted instruction performance of student pairs as related to individual differences. Journal of Educational Psychology, 1973, 65, 65-73.

THE PREDICTIVE VALIDITY OF OPERATIONAL AND EXPERIMENTAL VARIABLES FOR MENTAL GROUP IVS IN THE NAVY: A REVIEW AND SUMMARY OF THE FINDINGS FROM FOUR SETS OF "CULTURE FAIR" TESTS

Charles H. Cory

In the late 1960s and early 1970s a service-wide program, Project 100,000, was carried out, a major purpose of which was to lower the admission standards and to increase substantially the proportions of lower mental level personnel in the military services. Project 100,000 was based on a philosophy whose major tenants may be expressed as follows:

1. Lower aptitude personnel possessed positive traits conducive to effective on-job performance. However, these characteristics were not being tapped by the classification tests because the existing measuring instruments were developed to predict scholastic performance in Class A-schools.

2. Hence lower aptitude personnel could have performed adequately on-the-job in many situations where they were not being utilized.

3. Ergo, if appropriate tests were used, i.e., measures of attributes related to on-job performance and geared to the abilities of the low aptitude personnel, these personnel could be accepted in greater numbers and used more flexibly and profitably.

As a result of Project 100,000, the Navy undertook an extensive research effort to develop and validate personnel selection tests which would be predictive of the performance characteristics of low mental ability personnel. This report presents a summary of the results of the research.

PROCEDURE

For the studies, the comparison groups were defined on the basis of ranges

of scores on the Armed Forces Qualification Test (AFQT). Thus, the performance of Mental Group IVs (AFQT centile scores 10-30, inclusive) was compared with that of I-IIIs (AFQT centile scores 31-99), and further comparisons were made between the scores of Lo-IVs (AFQT scores 10-20) and those of Hi-IVs (AFQT scores 21-30).

Four separate sets of experimental tests were developed and validated on four separate samples of IVs and non-IVs. For each set, the testing was conducted early in recruit training and follow-up on-the-job was carried out a few months before the conclusion of the enlistment period of the Mental Group IVs. For Phases 2, 3, and 4, for which IVs had 2-year enlistment periods, the follow-up data were collected during the 21st or 22nd month of enlistment. For Phase 1, the follow-up was carried out during the 46th as well as the 21st month.

Dates of administration of the experimental tests and collection of the performance follow-up marks are shown below:

Sample	Period during which Experimental Tests were Administered	Date of Mailout for on-job performance follow-up
Phase 1	6/27/67 to 8/4/67	5/69 and 5/71
Phase 2	2/13/68 to 4/4/68	11/69
Phase 3	6/27/68 to 11/10/68	5/70
Phase 4	1/6/69 to 6/30/69	11/70

Experimental Tests

The experimental batteries were developed after an extensive literature

search for tests which had been useful or were thought to be useful for selecting low mental-ability personnel for job assignments. Nineteen tests were selected or developed. These included measures of cognitive skills, perception and coordination, biographical information, personality, and interests. Brief descriptions for each of the experimental tests together with a listing of the phase in which each test was used are shown in Table 1.

The cognitive skills which were tested, in the main, were those requiring "fluid" as opposed to "crystallized" abilities, as defined by Cattell (1957). Furthermore, the abilities in general were those required to deal with concrete in contrast to abstract problems or mental operations.

A variety of approaches were used to attempt to measure personality factors, interest, motivation, and biographical variables both in terms of self report and measurement in situ. Vocational interests were measured by the Strong Vocational Interest Blank (SVIB) and an interest measure developed on Navy job elements. Two specially constructed inventories, the Performance Index and the Recruit Attitude Questionnaire, were used to measure personal interests and attitudes. Background and developmental variables were measured by means of two biographical interest questionnaires each designed with particular emphasis on indices of socioeconomic status (SES), a variable which has frequently been found to be useful for identifying capable personnel. Physical skills and degree of motivation were measured by means of the Hand Skills and the Manual Speed Tests. Coordination, integrative ability and foresight were tested using an adaptation of the Porteus Maze Test. Self report measures of personality inadequacies and problems were secured by means of the Recruit Temperament survey--a test which was used operationally at the time to identify personality misfits for early discharge from the Navy.

Table 1

Descriptive Characteristics of the 19 Experimental Tests

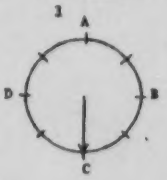
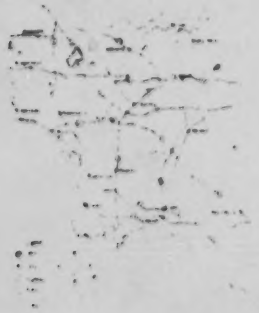
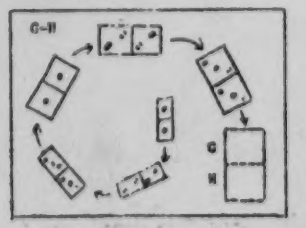
Test	Description	Illustrative Items	Phase
Cognitive Skills			
1. Memory for Numbers	A digit-span measure of immediate memory, administered by a tape recording. Spans range from 4 to 10 digits. A series of numbers is read aloud. After an 8-second pause subject records numbers remembered.	Script on tape: A 6-8-3-2; B 1-9-5-8-7; C 2-6-7-5-1; D 6-4-9-7-5-3. Answer Sheet: A [6][8][3][2]; B [1][9][5][8][7]; C [2][6][7][5][1]; D [6][4][9][7][5][3]	1
2. Readministration of the Armed Forces Qualification Test (AFQT)	The AFQT is a selection test used for the Armed Services. It contains four 25-item subtests measuring the following areas: (1) Vocabulary and Verbal Reasoning (2) Knowledge of Tools and Mechanical Parts, (3) Arithmetic Reasoning, and (4) Spatial Reasoning	Since the content of AFQT is classified restricted, illustrative items from the test it cannot be shown.	1
3. Listening Skills 1	A tape-recorded test involving oral directions in relation to graphic stimuli presented in an accompanying test booklet.	Script on tape: In your imagination turn the arrow 3/4 of A turn clockwise. Then 1/4 of a turn counterclockwise. Then turn it 6 points clockwise. Where will the arrow be pointing?, Test booklet: 1 	
4. Listening Skills 2	Similar to LS-1	Script on tape: Begin at the Airfield. (pause) Go North to Highway 2 and turn West. Go as far as the next highway and turn North. What is the first town you will come to? Test booklet: 3 	
5. Dominoes	Pictorial problems of dominoes arranged in patterns and series	Decide how many dots should be on each half of the last domino. There can be from 0 to 6 dots on each half. Write the correct number of dots for each half on your answer sheet. 	2

Table 1 (contd)

Descriptive Characteristics of the 19 Experimental Tests

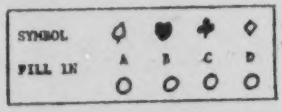
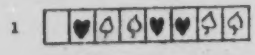
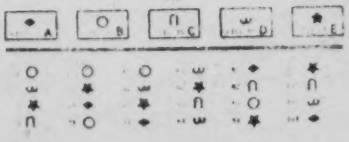
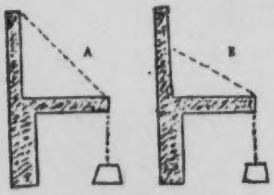
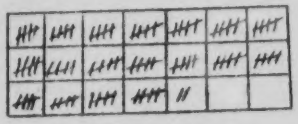
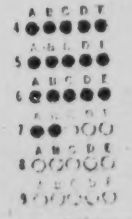
Test	Description	Illustrative Items	Ph
6. Card Pattern (CARD)	Pictorial problems using playing card symbols arranged in patterns and series.	Choose the symbol that belongs in the empty box. Use the answer code to find which circle to fill in on your answer sheet. Answer Code  1  2	4
7. Coding	A symbol-substitution test involving five figures corresponding to response circles on the answer sheet.	Mark the letter on your answer sheet that stands for each figure shown. Use the answer code given below 	4
8. Mechanical Principles	Pictorial problems involving applications of physical and mechanical principles and relationships.	 Which chain will break first?	4
Motivation and Perceptual Speed			
1. Hand Skills	Answer sheet contains section of numbered boxes. The task is to make a series of five tally marks (viz., 𝍸) in each box as rapidly as possible. At the beginning of each separately timed section, a "passing" score is announced. The number of tallies completed beyond the "passing" score is considered a measure of motivation.	Answer Sheet 	1
2. Manual Speed	Developed as an electronically scoreable form of the Hand Skills Test. The task is to fill in as many circles on an machine-scorable answer sheet as possible in the time limits.		1

Table 1 (contd)

Descriptive Characteristics of the 19 Experimental Tests

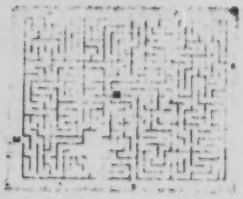
Test	Description	Illustrative Items	Phase
3. Maze Test	A group-administered true-false test given under speed conditions and patterned after the Porteus Mazes.	For every door of each maze, mark Circle A or B to show if the path to the center is open or blocked. You may mark the path with your pencil, if you wish. 	2
4. Word Finding	Items involve matching stimulus and response words from two columns. Initially, words are in close vertical proximity and gradually they become more distant.	For every word on the left find the answer word in the right column. On your answer sheet, Mark the letter printed after the answer word. 1 DRIVE — game-D, drive-A 2 FORK — heater-D, suit-E 3 STREET — button-D, street-E 4 SUIT — berry-B, ribbon-A	4
Vocational Interest 5			
1. Job Check list	Designed as a vocational interest inventory; uses descriptions from actual jobs performed by enlisted men in the military. Yields scores on 7 scales derived by cluster analysis.	Indicate Like, Dislike or Indifferent to each activity. 1. Use a file or grindstone to smooth edges of metal parts. 2. Use an adding machine to check payroll records. 3. Repair refrigerators and air conditioners.	2
2. Strong Vocational Interest Blank	A commercially-distributed inventory of compatibility of individual interests to those of successful employees in over 50 occupations. (A glossary was developed, on the basis of trial administration to recruits, which provided definitions for the terms which were unfamiliar.)	Indicate whether or not you like the following kinds of work 1. Actor 2. Editor 3. Farmer	3
3. Performance Index	Designed to provide an inventory of work values and characteristics and attitudes toward jobs, other individuals, and social concepts.	Indicate if True or False for you: 1. I like to bluff when I get in a tight spot. 2. I can get others interested in working with me Which of the following two kinds of work do you prefer? 1. Deal with things 2. Deal with people Indicate how well each of these items describes you: 1. Pay attention to details. 2. When caught in a mistake, I make excuses.	3

Table 1 (contd)

Test	Description	Illustrative Items	Phase
Experiences, Interests and Attitudes			
1. Biographical Information Form	Designed to tap education and work experiences, socioeconomic status, self-image, attitudes toward parents and society, and plans for the future.	Which of the following jobs have you had and how did you like it? 1. Paper route. 2. Farm or ranch hand. Which of these things have you done? 1. Drove a truck. 2. Operated power tools. Agree-Disagree: 1. Good luck is more important than hard work for success. 2. I sometimes feel I just can't learn. How do you compare with other men your age on-- 1. Doing hard physical work. 2. Making decisions.	2
2. Recruit Attitude Questionnaire	A paired-comparison test which measures attitudes toward nine aspects of enlisted service in the Navy.	For each of the following questions, record which of the two values is more important to you. 1.--Travel and adventure. --Education and Job Training. 2.--Family Life. --Patriotism and Duty.	3
3. Biographical Information Questionnaire	Designed to assess cultural background, exposure, and socioeconomic status as reflected by objects in the home, activities of a recreational or social nature, and biographical data.	1. What kind of neighborhood did you live in when you were a teenager? 2. Where did your money come from when you were of high school age? How often did you do the following activities? 1. Make minor houshold repairs 2. Check out a library book	4
4. Recruit Temperament Survey	True-false questions concerning attitudes, interests, physical characteristics, outlook on life, and physical complaints.	1. Do other people often take advantage of you ? 2. Are you good at handling younger children? 3. Do you often have trouble getting to sleep?	4

The experimental tests were developed to require little reading, and the level of the reading material which was included was at or below the fifth grade. In addition, the difficulty levels of the items in the experimental tests were relaxed so that the means of the tests were about 70% of the maximums. This was about 10 to 15 percentage points higher than the means of the operational tests.

A brief description of the variables derived from operational records, either classification tests or biographical information which were used for the study is given in Table 2.

Test-Retest Reliabilities of the Cognitive Experimental Tests

Test-retest reliabilities of the major cognitive experimental tests were computed on separate samples of recruits at NTC, San Diego. For this part of the study, the tests were administered in June 1971 and the retests were carried out about five weeks later. The sample sizes and the reliabilities are shown in Table 3.

Most of the reliabilites of the experimental tests were in the 60s and 70s, but two, Memory for Numbers and Card Pattern, had reliabilities in the 30s.

The major practical consequence of the relatively low reliabilities of the tests in the experimental batteries was that they slightly penalized the validity coefficients obtained for the experimental tests in relation to those obtained for the operational tests (which had test-retest reliabilities in the 70s and 80s).[1]

After the test scores were processed and follow-up data were collected, the data were further subgrouped into Apprenticeshp and Technical ratings. The

---

[1] Swanson, L., NPRDC personal communication.

Table 2

Operational Variables Used in the Study

Variable	Acronym	Response Mode	Content Areas
		Selection and Classification Tests:	
Armed Forces Qualification Test	AFQT	Multiple Choice	Vocabulary, Arithmetic reasoning, spatial reasoning, and numerical knowledge.
General Classification Test	GCT	Multiple Choice	Work meanings and the ability to reason verbally.
Arithmetic Reasoning Test	ARI	Multiple Choice	Quantitative aptitude, including mathematical reasoning and problem solving.
Mechanical Test	MECH	Multiple Choice	Basic mechanical and electrical knowledge and mechanical principles.
Clerical Test	CLER	Multiple Choice	Perceptual speed and accuracy.
Sonar Pitch Memory Test	SONR	Multiple Choice	Ability to perceive and remember small differences in tonal pitches.
Radio Code Aptitude	RADO	Multiple Choice	Ability to learn, remember and use sound patterns as symbols.
Electronic Technician Selection Test	ETST	Multiple Choice	Mathematics, science, electricity and electronics.
Shop Practices	SHOP	Multiple Choice	Knowledge of tools and shop equipment.
Year of Birth	YRBI	Free Answer	
Year of Enlistment	YREN	Free Answer	
Years of Education	ED	Free Answer	
High school graduate vs. non-High school graduate	HS	Binary	High school graduate = 1 Non-graduate = 0
Black vs. non-Black	BL	Binary	Black coded 1, Non-Black coded 0
Lo-IV vs. Hi-IV	LA	Binary	Personnel with AFQT scores 10-20 coded 1 (Lo-IVs), personnel with AFQT scores 21-30 coded 0 (Hi-IVs).

TABLE 3

TEST-RETEST RELIABILITIES OF THE COGNITIVE EXPERIMENTAL TESTS

TEST	R_{++}	N
MEMORY FOR NUMBERS	.34	110
LISTENING SKILLS 2	.69	129
DOMINOES	.66	172
CARD PATTERN	.33	177
CODING	.47	141
HAND SKILLS TEST, TOTAL TALLIES IN PART 3 MINUS THE BASE RATE	.67	172
HAND SKILLS TEST, TOTAL TALLIES IN PART 4 MINUS THE BASE RATE	.71	172
HAND SKILLS TEST, TOTAL TALLIES	.68	172
MANUAL SPEED TEST TOTAL MARKS	.57	133
MAZE TEST	.77	149
WORD FINDING	.66	143

821

Apprenticeship group consisted of nonrated, nonstriker personnel in the Seaman, Fireman, Constructionman and Airman ratings. The Technical group consisted of designated strikers and rated personnel in all but Apprenticeship ratings. Thus, the Apprenticeship group included personnel in pay grades E1 through E3 and the Technical Group included personnel in pay grades E3 through E5.

Criteria

The operational and experimental tests were evaluated in terms of their ability to predict three criteria: on-job performance, ability of IVs to advance into a Technical rating, and retention of IVs throughout their period of enlistment.

The on-job performance criterion consisted of a special supervisory rating collected by mailout of the performance of each man in relation to that of other personnel in the same rate. It was measured on a 5-point scale in which code 1 indicated performance in the bottom 20%, code 2 indicated performance in the bottom 21-40%, etc.

The advancement criterion was a binary variable which was formed for the IVs originally tested who were found on the Enlisted Master Tape Record at the time of the follow-up. Mental Group IVs who were found to be assigned to a Technical rating at that time were coded 1 and those who were assigned to an Apprenticeship rating were coded 0.

The retention criterion was a binary variable which was formed for IVs in the Phase 4 sample. The IVs who were from the original sample who were not in the Navy at the time of the follow-up were coded 0 and IVs who remained in the Navy at that time were coded 1.

Analyses

Data for the four samples were analyzed separately.

The ratings into which IVs were found to have been appointed were identified and classified in terms of their percentages of IVs. The mean overall on-job performance marks of IVs and non-IVs were compared. Zero-order and multiple regression validity coefficients between the experimental and operational variables and the criteria were computed for each of the samples.

RESULTS

Ratings to Which IVs Were Assigned

Over the four studies, IVs were found in 51 Technical and 4 Apprenticeship ratings. About 20% of the IVs present at the time of the follow-up were assigned to Technical ratings with by far the greatest proportion of IVs in the Technical group being Stewards.

Except for Stewards, there was very little concentration of IVs in particular Technical ratings. In only two other ratings, Equipment Operator and Commissaryman, did IVs constitute 5% or more of IVs in the Technical group. The IVs in each of an additional seven ratings, Hull Technician, Engineman, Boatswain's Mate, Hospital Corpsman, Ship's Serviceman, Yeoman, and Machinist's Mate constituted between 3 and 5% of IVs in the Technical group.

Relative Performance of IVs and non-IVs

For both Apprenticeship and Technical groups, mean on-job performance of IVs for all four phases of the research (Table 4) was lower than that of I-IIIs. For the Apprenticeship group the IV deficits were statistically significant for three of the four studies, whereas the deficits of IVs in the Technical group were generally not significant. This is another instance (see Cory 1976) in which IVs in Technical ratings were found to have consistent but relatively small deficits in performance in comparison with I-IIIs.

TABLE 4

DESCRIPTIVE STATISTICS OF SUPERVISORS EVALUATION OF ON-JOB PERFORMANCE OF NON-IVS, IVS AND HI- AND LO-IVS IN FOUR SAMPLES

	APPRENTICESHIP			TECHNICAL		
	$\bar{X}_{N\text{-}IV}$	$\bar{X}_{IV}$	R_{L4}	$\bar{X}_{N\text{-}IV}$	$\bar{X}_{IV}$	R_{L4}
PHASE 1						
1ST FOLLOW-UP	2.64	3.21	-.01	3.69	3.54	-.03
2ND FOLLOW-UP	3.54	3.32	-.19	3.78	3.57	-.01
PHASE 2	3.31**	2.94	-.05	3.69	3.67	-.26
PHASE 3	3.48**	2.84	-.01	3.73	3.65	-.12
PHASE 4	3.56**	3.20	-.08	3.77*	3.32	-.09

NOTE:

1. STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE CELLS MEANS OF IVS AND NON-IVS ARE INDICATED BY ASTERISKS BETWEEN THE TWO CELL MEANS.
2. NONE OF THE R_{L4} VALUES WAS STATISTICALLY SIGNIFICANT.

*$p < .05$

**$p < .01$

As shown by the validity coefficient of L4, Lo-IVs had consistent non-significant performance deficits in relation to Hi-IVs.

Validities of Operational Variables for On-Job Performance

In a preliminary analysis of the validity data substantial numbers of significant validity coefficients were found for every sample except Phase 1. Despite two performance follow-ups for Phase 1, a total of only two coefficients were statistically significant for IVs and 11 coefficients were statistically significant for non-IVs. In particular the very low incidence of significant validities for the classification tests and education variables in Phase 1 contrasted with the relatively consistent patterns of significant validities of these variables in the other three studies. This suggested that the Phase 1 data were biased by abnormal influences. Therefore, consideration of the validites of the operational tests for on-job performance was narrowed to include only those in the Phases 2, 3, and 4 samples, Table 5 presents these relationships. It gives the total number of statistically significant validity coefficients for each variable in the three phases. Of course, 3 is the maximum number for any cell in the N column of the table.

Apprenticeship Group

For the Apprenticeship group about the same percentage of operational variables had statistically significant validities for IVs as for I-IIIs, but the types of variables which were significantly valid differed markedly for the two mental levels. Thus, for IVs, MECH and RADO had significant validities for all three samples and CLER, SHOP, and High School Graduation were also consistently significantly valid.

The pattern of significant validities of I-IIIs differed most markedly from that of IVs in the following respects: (1) ARI and GCT were somewhat better

TABLE 5

INCIDENCE IN PHASES 2, 3 AND 4 OF STATISTICALLY SIGNIFICANT ZERO-ORDER VALIDITIES FOR ON-JOB PERFORMANCE

	I-III				IV			
	APPRENTICESHIP		TECHNICAL		APPRENTICESHIP		TECHNICAL	
PREDICTOR VARIABLE	N	%	N	%	N	%	N	%
AFQT	2	67	2	67	1	33	1	33
GCT	2	67	3	100	1	33	0	0
ARI	3	100	3	100	1	33	0	0
MECH	2	67	3	100	3	100	0	0
CLER	2	67	3	100	2	67	1	33
SONR	1	33	0	0	1	33	0	0
RADO	1	33	1	33	3	100	0	0
ETST	1	33	3	100	0	0	0	0
SHOP	1	33	3	100	2	67	0	0
YEAR OF BIRTH	0	0	2	100	2	100	1	50
HIGH SCHOOL GRADUATE	2	67	2	67	2	67	2	67
BLACK/NON-BLACK	0	0	0	0	1	33	0	0
YEARS OF EDUCATION	3	100	3	100	1	33	2	67
LO-IV/HI-IV	-		-		0	0	0	0
TOTAL/AVERAGE	20	53	28	74	20	49	7	17

as predictors and MECH, RADO, and SHOP were somewhat worse; (2) ED was better than HS as a predictor for the non-IVs. For IVs the obverse was true.

Technical Group

Even greater difference existed between IVs and non-IVs in the Technical group in the number and types of variables having significant validities for on-job performance. Only two classification tests were valid for IVs, and both High School Graduation and Years of Education were better for IVs in the Technical group than any of the classification tests. In contrast, for the non-IVs, 19 of the 24 validity coefficients for the operational tests were statistically significant. For these personnel all of the Navy classification tests except SONR and RADO were statistically significant for all three phases.

Validities of Experimental Tests for On-Job Performance

Zero-Order Validities. Validities for experimental variables for which either IVs or non-IVs had statistically significant validities are shown in Table 6. A variable was included in Table 6 if the validity of the variable for either IVs or non-IVs was statistically significant for on-job performance. If a variable was included, the validity coefficient for the variable for the other mental group was also included.

The percentage of significant validity coefficients for I-IIIs was greater than that for IVs. In Apprenticeship ratings 15 of the validity coefficients of I-IIIs (17%) were statistically significant compared with 9 of the coefficients for IVs (10%). For the Technical group these percentages were 36 and 14 for the I-IIIs and the IVs, respectively.

IVs in the Apprenticeship group with high on-job performance were significantly predicted by the following characteristics: (1) admission of the IV that he has little chance in life, (2) raised in a 2-parent household, (3) self reported leadership and interpersonal relations skills, (4) self reported interests

Table 6

Significant Zero-order Validities of Experimental Variables for IVs and Non-IVs, in any of the Four Research Phases

Experimental variable	Validity for Global Performance Marks							
	Apprenticeship Group				Technical Group			
	IV		Non-IV		IV		Non-IV	
	r	N	r	N	r	N	r	N
Phase 1								
Digit Span, First Error Span	.21	53	.32*	39	.32*	40	.02	523
Hand Skills Test, Total tallies in Part 3 minus the base rate	.06	114	.17*	102	.01	103	.07*	1214
Hand Skills Test, Total tallies in Part 4 minus the base rate	.04	114	.25*	102	-.03	101	.10**	1214
Hand Skills Test, Total tallies	.03	114	.25*	102	.00	103	.09**	1214
AFQT Math Subtest					.10	107	.07*	1255
Manual Speed, Total					.18	77	.06*	1139
Phase 2								
Dominoes					.04	19	.08**	1191
BIF, Father's level of education					.41	22	-.09**	1788
Parental cohabitation	-.15*	260	-.11*	369				
Number of siblings who went to college					-.32	26	.07**	1771
Level of educational attainment expected of individual by parents					.04	26	.09**	1787
Opinion that individual has little chance of success in life.					.09	26	.06*	1826
Retention scale	.13*	220	.00	525	.01	43	.08**	2461
	.11	424	.10*	781				
Job Check List Scales								
Clerical Interest	-.06	349	-.10*	779				
Domestic Activities Interest					-.03	36	.06**	2461
Graphics Interest					-.39*	36	-.01	2461
Clerical Hand Skills Interest					-.37*	36	-.05*	2461
Phase 3								
RAQ Scales								
Fringe Benefits & Retirement					-.01	18	.21**	277
Equipment & Facilities	.06	34	.16*	193	.25	18	.21**	277
Patriotism & Duty	.34*	34	.06	193				
Pay					-.02	18	-.15*	277
Advancement	-.23	34	-.22**	193				
SVIB Scales								
Business Management					.11	127	.07*	1686
Office Practices					.18*	127	.07*	1686
Merchandising					.19*	127	.08**	1686
Technical Supervision					.17	127	.10**	1686
Mathematics					.12	127	.10**	1686
Mechanical					.17	127	.07*	1686
Navy Interest					.03	127	.09**	1686
Liberal/Conservative					-.20*	127	-.10**	1686
Masculinity/Femininity					.01	127	.09**	1686
Phase 4								
Coding	.14**	1664	.11*	618	.28**	98	.11**	1629
Word Finding	.15**	1665	.13*	618	.14	98	.09**	1629
Mechanical Principles					.50**	24	.05	740
BIQ Scales								
Mechanical Activity	.14*	907	-.08	422				
Economics	.07	907	-.07	422	.33*	35	.04	960
Cultural Activities	.07	907	-.13*	422				
Interpersonal Interaction	.16**	907	-.05	422	.06	35	.07*	960
Self perceived socio-economic status	.04	907	.13*	422				
Factorically determined bio scales								
Activities and Experiences					.50**	32	.03	909
Electro/Mechanical activities	-.02	686	-.27**	368	-.38**	32	.07*	909
Socio economic status					.15	32	.07*	909
Social Leadership	.35*	686	-.06	368	.19	32	-.09*	909
Recruit Temperament Survey	-.11**	702	-.16*	181	-.32**	61	-.05	539

* $p<.05$
** $p<.01$

and past experience in mechanical types of activities, and (5) relative freedom from fears and neuroses.

On the other hand the better Technical IVs tended to have high scores on Coding, Mechanical Principles, and Digit Span, to have low interests in graphics and hand skills, to have high interests in Office Practices, Merchandising, and Economics, and to be generally conservative in outlook. In addition, they had greater experience in electro-mechanical activities, a wider and more diversified social and educational background and more balanced, less neurotic personalities.

Multiple Regression Validities

Multiple regression analyses using a predictor selection technique, accretion paradigm, produced maximum shrunken validities ranging from 0 for the Phase 1 sample to .39 for the sample for Phase 4. The average maximum shrunken validity for the last three phases was about .30. In general, operational tests were selected first by the program and about 20 points of the validity were attributable to them. From four to ten points were attributable to biographical variables and education and experimental variables.

For the shrunken validities in the Technical group the operational tests contributed about 35 points to the shrunken multiple validity coefficient of the last two phases and education and experimental variables added about 15 points to the validity coefficients, on the average.

Significant Predictors of IVs Who Were Able to Advance into Technical Ratings

Table 7 shows the validity coefficients for advancement of the operational variables for IVs in Phases 1 through 4. Table 8 shows similar information for the experimental variables which had statistically significant validities.

For advancement as was the case for on-job performance, most of the

TABLE 7

ZERO-ORDER VALIDITIES OF OPERATIONALLY - DERIVED VARIABLES FOR IVS FOR THE BINARY ADVANCEMENT CODE

	R			
PREDICTOR VARIABLE	PHASE 1	PHASE 2	PHASE 3	PHASE 4
AFQT	.22**	.11**	.37**	.14**
GCT	.28**	.23**	.01	.26**
ARI	.18**	.27**	.10*	.17**
MECH	.06	.00	-.14**	.07*
CLER	.19**	.16**	.22**	.12**
SONR	.16**	.07*	-.03	.08**
RADO	.20**	.16**	.08	.08**
ETST	.18**	.21**	.02	.12**
SHOP	.18**	.10**	-.11*	.12**
YEAR OF BIRTH (YRBI)	-.05	-.10**	--	.06*
HIGH SCHOOL (HS)	.24**	.12**	.22**	.04
BLACK/NON-BLACK (BL)	.01	.06	-.26**	.11**
YEARS OF EDUCATION (YRED)	.23**	.43**	.15**	.07*
LO-IV/HI-IV (L4)	-.16**	.09**	.24**	.11**

*P<.05; **P<.01

Table 8

Experimental Variables Having Significant Zero-order Validities for the 1st Advancement

Experimental Variable	Validity for Global Performance Marks	N
Phase 1		
Digit Span, Total Correct	.27**	290
Digit Span, Weighted Total	.23**	358
Manual Speed Test, Difference Score	.18**	340
Manual Speed Test, Total	.14**	339
Hand Skills Test, Part 4 minus Base Rate	.25**	339
Hand Skills Test, Total Score	.24**	349
Verbal Section, AFQT Retest	.20**	363
Mathematics Section, AFQT Retest	.21**	363
Phase 2		
Dominoes	.20**	582
Maze	.17**	252
Listening Skills Test, Form 1	.17**	283
BIF Questions		
Family financial position in comparison with classmates	-.10**	562
Number of siblings who went to college	.10**	514
Job Check List Scales		
Mechanical Interest	.11**	722
Communications Interest	.08*	722
Verbal Activities Interest	.09*	722
Construction and Fabrication Interest	.11**	722
Heavy Equipment Operator Interest	.10**	722
Clerical Hand Skills Interst	.11**	722
Phase 3		
Recruit Attitude Questionnaire Scales		
Patriotism and Duty	.32*	54
Travel/Adventure	-.29*	54
Pay	-.34*	54
Advancement	.39**	54
SVIB Scales		
Business Management	.32**	377
Merchandising	.34**	377
Office Practices	.42**	377
Military	.47**	377
Technical Supervision	.19**	377
Mathematics	.58**	377
Science	.43**	377
Mechanical	.30**	377
Nature	.39**	377
Teaching	.35**	377
Navy Interest	.14**	377
Achievement	.55**	377
Liberal/Conservative	.25**	377
Masculinity/Femininity	.11*	377
Social Introversion	.37**	377
Occupational Level	.15**	377
Phase 4		
Coding	.08**	1,889
Card Pattern	.07*	1,145
Word Finding	.11**	1,890
BIQ Scales		
Domestic	.14**	1,011
Recreational Activities	.09**	1,011
Economic	.14**	1,011
Interpersonal Interaction	.08**	1,011
Socio-economic Scale	.09**	1,011
Self-perceived Social Class	.08*	817
Overall Activity	.11**	1,011
Social Leadership (factorially derived)	.10**	775

*p<.05
**p<.01

operational tests had statistically significant validity coefficients across the samples. Thus, AFQT, CLER, SHOP, Years of Education, GCT, SONR, RADO, and High School graduation had significant validity coefficients for either three or all four of the samples. The validity coefficient for BL indicates that there were no differences in the proportions of Black and non-Black IVs who advanced into Technical ratings.

Surprisingly, the validity coefficients of the operational variables for advancement were generally higher than those for these variables for on-job performance. This is despite the truncation of the effective maximum validity for a binary criterion in contrast to the maximum for a criterion with a 5-point scale.

In general a considerably larger proportion of experimental tests were predictive of advancement than of on-job performance of IVs. Whereas only 12% of the validity coefficients for the experimental tests were statistically significant for on-job performance of IVs, 59% of the coefficients for those variables were significant for advancement. Furthermore, the significant validity coefficients for advancement were generally higher than those for on-job performance.

By all odds, scales from the Strong Vocational Interest Blank (SVIB) had the highest validities for advancement of any variables in the study. The SVIB scales having the highest validities were Military, Mathematics, Science, Office Practices, and Achievement.

The average shrunken multiple R for the operational tests for the four Phase samples was about .30. Education and biographical variables when added to the battery increased the shrunken validity coefficient by eight points, on the average. Experimental tests when added to the battery of classification tests and operational variables increased the shrunken R two points for the Phase 1 sample

and 21 points for Phase 3, the sample to which the SVIB was administered. Operational tests were the first variables selected by the multiple regression program for all samples except Phase 3. As would be expected for that sample, SVIB scales were selected first.

Significant Predictors of IVs Who Were Retained Throughout Their Enlistment Period

The two variables which were most predictive of retention of IVs were AFQT score and the binary L4 code, with zero-order validities of .50 and .57, respectively. No other experimental or operational variable had validity greater than .19. Multiple regression analysis produced a shrunken multiple R of .58 with all but one point of the R being predicted by L4. For this sample it was found 2% of the Lo-IVs had been retained throughout their enlistment period, compared with 50% retention for the Hi-IVs.

SUMMARY AND COMMENTS

The findings of the study were consistent with those of previous Navy studies in indicating that a percentage of the IVs selected can perform about as well as I-IIIs. However, they indicate that the percentage is likely to be quite small. The present study found only about 20% of the IVs remaining in the Navy at the conclusion of their enlistment period to be in Technical ratings. In addition, IVs were concentrated in a very narrow band of Technical ratings, with the only rating in which they had substantial numbers being Steward and the only other Technical ratings which had some concentration of IVs being Equipment Operator, Commissaryman, Hull Maintenance Technician, Engineman, Boatswain's Mate, Hospital Corpsman, Ship's Serviceman, Yeoman, and Machinist's Mate.

The findings were also consistent with those of other Navy studies in that the deficits found between the mean on-job performance marks of IVs in Technical

ratings and those of I-IIIs in those ratings were usually nonsignificant. On the other hand the substantial and significant deficits in on-job performance of IVs in the Apprenticeship group in relation to I-IIIs in that group seems somewhat surprising. The simple and undemanding nature of the jobs in the Apprenticeship group would suggest that IVs would be about as well suited for them as I-IIIs.

The findings did not support the point of view that culture-fair tests would provide a better means of selecting IVs that the operational classification tests. In particular the figures suggest that few if any of the Mental Group IVs in the Lo-IV range (AFQT scores from 10 to 20) are suitable for the Navy.

Comparisons concerning the relative predictiveness of the operational and experimental tests may be looked at in two ways, On the one hand, the experimental tests represent an approach which largely did not work. These intended culture-fair tests simply were not as effective for identifying low mental ability personnel who could perform well on the job as were tests in the Navy classification battery. On the other hand, it is gratifying to know that the Navy classification tests, which had been developed to predict academic criteria (grades in Class A School) were also effective as predictors of on-job performance not only in Technical ratings, but also in Apprenticeship ratings.

For Mental Group IVs who apply for enlistment in the Navy in the future, the findings suggest that classification tests together with scales from the SVIB or some other measure of vocational interests can be used to select with considerable accuracy the IVs who have the most potential for advancing into Technical ratings.

REFERENCES

Cattell, R. B. Personality & Motivation; Structure & Measurement. Yonkers on Hudson, New York: World Book Company, 1957, p. 618-19

Cory, C. H. A Comparison of the Job Performance and Attitudes of Category IVs and I-IIIs in 16 Navy Ratings (NPRDC Tech. Rep. 76-35). San Diego, CA: Navy Personnel Research and Development Center, May 1976 (NTIS No. AD-A024-642)

ESTABLISHING TRAINING PRIORITIES IN AN ARMY SUPPLY MILITARY OCCUPATIONAL SPECIALTY

Raymond O. Waldkoetter, Ed.D.
Arthur C. F. Gilbert, Ph.D.
James L. Raney, Ph.D.

U.S. Army Research Institute for the Behavioral and Social Sciences[1]
Arlington, Virginia 22209

Helmut H. Hawkins

Plans and Occupational Development Division
U.S. Army Military Personnel Center
Alexandria, Virginia 22331

Establishing training priorities for the different elements of instruction prerequisite to job performance becomes a pivotal element in curriculum design. The end result is the identification of those tasks that should be included or excluded from the course curricula and answers the question of "Where should this task be taught?". An even more fundamental question to which such a system can provide an answer is, "Is this knowledge necessary to perform the job?" If the answer is affirmative then an adequate system of determining training priorities will yield an answer to the question of "When should this knowledge be imparted?". Finally, such a system can resolve the question of "To what level should the knowledge be imparted?". Thus, from a training viewpoint an adequate system of training priorities will determine if, and to what level, when and where the knowledge necessary for job performance should be imparted. Technical training objectives can then be adequately assessed by such a method.

Training priorities can be analyzed to define job dimensions that lead to delineating personnel qualifications, skills and experience necessary to perform the job. In this instance the training priority system becomes basic to the selection and classification system.

Again, the training priority system has another use that is of paramount importance in the military-industrial setting of becoming the basis for proficiency testing and performance evaluation measures tailored to the job being performed.

---

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research Institute, the Military Personnel Center, or the Department of the Army.

Finally, another but a highly relevant issue can be accommodated and this is the one of job enrichment. If the incumbent in a particular position is not afforded the opportunity to perform those tasks for which he is trained, job dissatisfaction may result. Therefore, the concept of training priorities can merge to determine at how near the point of application training should be given.

The concept of training priorities model employed in this research was based on the concept of task criticality developed in the United States Air Force. Results of recent research (Meade, 1975) indicate the potential utility of a four factor or four dimensional model that is highly correlated with the training priorities that should be assigned to contain tasks within the context of overall job performance. These four factors were task learning difficulty (TLD), consequences of inadequate performance (CIP), task delay tolerance (TDT), and percent of members (PCT) in the particular military occupational specialty who perform the specific tasks.

The results of the research performed in the Air Force (Meade, 1975) raised the question of the applicability and utility of the Air Force developed methodology to determining task training priorities in an Army setting. Consequently, the purpose of this research was to determine the extent to which the four factor model (i.e. TLD, TDT, CIP, and PCT) was capable of predicting task training priorities. A further innovation was to determine whether the four factors could be used to predict judgments of specific kinds of training.

PROCEDURE

As a result of an operational requirement the MOS selected to test the utility of the four factor model was 76V Equipment Storage/Specialist. A task list and rating scales were developed and administered to a representative sample of supervisors and instructors in this MOS.

Task List

The task list consisted of 183 tasks extracted from the comprehensive Occupational Data Analysis Program (CODAP) survey booklets. The task list contained only those tasks that were performed by men in grades E-1 through E-5 and excluded those duties that were strictly military in nature.

Task Rating Scales

Four rating scales were developed. One rating scale was the training priority criterion measure while the other three involved measurement of those variables that were considered to be predictive of the training priority evaluation (i.e. TLD, CIP, TDT).

Type of Training Scale

The type of training scale or the criterion measure required the rater to evaluate the 183 tasks on a five point scale as to the type of training that he considered to be the most appropriate for each task. The rating categories were as follows: (1) no training required (2) on-the-job training (3) formal unit training (4) non-resident school training (5) resident school training. The raters were informed that the scale was a measure of the most appropriate type of training for the successful performance of a task and that only one of the five options was to be selected which in the rater's judgment appeared to be the best method of training for that task.

Task Learning Difficulty

This scale was defined as "a measure of the need for lengthy systematic training before a new member of the appropriate Army specialty could perform the task adequately". It was further defined in terms of the "difficulty involved in "picking-up" the task on the job without any systematic training". This scale was also a seven-point scale.

Consequences of Inadequate Performance

This scale was defined as a measure of the probable consequences of inadequate performance of a task. It was stated in the instructions that such consequences could be possible injury or death, wasted supplies, damaged equipment, or wasted manhours of work. A seven point scale again was used. The scale value of "1" corresponded to the least serious consequences of inadequate performance defined as "Extremely low--if the task is not done correctly, the possible consequences of inadequate performance are negligible". The high end of the scale, corresponding to a scale value of "7", was defined as "Extremely high--if the task is not done correctly, the possible consequences of inadequate performance are disastrous".

Task Delay Tolerance

This scale was defined as a measure of how much delay of task performance can be tolerated between the time the soldier becomes aware that the task must be performed and the time at which he must start doing it. In other words, does he have to perform the task immediately or does he have time to call his supervisor, read a field or technical manual, or even will the time permit his being actually taught how to perform the particular task. The ratings were obtained on a seven-point scale where the scale value of "1" was defined as "Extremely low--the task must be performed immediately whenever it is encountered". The scale value of "7" was defined as "Extremely high--task performance is almost never urgent".

Subjects

Subjects were 80 non-commissioned officers in MOS 76 V Equipment Storage Specialist. These non-commissioned officers were holding positions as instructors or as supervisors. In the sampling plan an attempt was made to secure as representative a sample as possible both in terms of geographical location and type of unit (i.e. operational or support such as a supply depot). In Table 1 the location and number of subjects

Table 1

LOCATION AND GROUP MEMBERSHIP OF SUBJECTS CORRESPONDING TO ORDER OF PRESENTATION OF RATING SCALES

Location	Group Membership 1	2	3	4	5	Total
Ft Ord	2	1	1	1	1	6
Ft Hood	2	2	2	2	2	10
Ft Devens	0	2	1	1	0	4
Ft Lee	5	5	5	5	5	25
Letter Kenny Army Depot	2	1	1	1	2	7
Ft McClellan	1	1	2	2	1	7
Ft Campbell	2	2	2	1	2	9
Ft Jackson	1	2	2	2	1	8
Ft Bragg	1	1	1	1	0	4
Totals	16	17	17	16	14	80

at the various locations are shown. The sample did not include subjects at overseas locations because of the operational requirements to complete the data collection and preliminary analyses at an early date.

Administration

Subjects were divided into five groups on the basis of order of administration of the scales. In four of the groups the orders of presentation of the scales are as shown in Table 2. Subjects were required to rate the tasks on the four scales in the order shown and were not allowed to refer to the ratings given to a task on a scale completed earlier than the one on which they were then completing. In the fifth group all of the scales were presented simultaneously and subjects were required to rate each task on all of the four scales before proceeding to the next. After completion of the rating tasks

Table 2

ORDER OF PRESENTATION OF THE RATING SCALES TO EACH OF THE FIVE GROUPS OF SUBJECTS

Group	Order of Presentation			
	1	2	3	4
1	TOT	TLD	TDT	CIP
2	TLD	TDT	CIP	TOT
3	TDT	CIP	TOT	TLD
4	CIP	TOT	TLD	TDT
5	Each task rated on scales at the same time			

Note: TOT = Type of Training
TLD = Task Learning Difficulty
TDT = Task Delay Tolerance
CIP = Consequences of Inadequate Performance

subjects were required to complete a questionnaire which contained items dealing with training and personnel management.

After completion of the rating tasks and questionnaire the subjects were interviewed either individually or in groups. The interview was unstructured but focused on issues such as the adequacy of the task list, problems encountered in the rating task, personnel utilization practices and possible improvement of training.

RESULTS

The data were first analyzed to determine the reliability or the amount of agreement among the raters for each of the three scales. These reliabilities were computed using the data from all of the 80 subjects without the deletion of possible erroneous ratings.

The reliability estimate for the Task Learning Difficulty Scale was .88 which was the highest reliability estimate obtained. Reliability estimates of .59 and .50 were obtained for the scales designated as Consequences of Inadequate Performance and Task Delay Tolerance respectively. A possible explanation of the relatively low reliabilities of these two scales is that the instructions may have been somewhat ambiguous. Also, it should be noted that the Task Delay Tolerance Scale was a reversed scale in comparison with the other two scales. In other words, the value of "1" on this scale was "Extremely low--the task must be performed immediately whenever it is encountered". Thus the relatively high importance for a task on this scale had a value of "1" whereas the relatively high importance had a value of "7" on the other two scales.

After the reliability estimates among raters for each of the four scales were established the mean rating that each task was given by the raters was computed for each of the four scales. This was accomplished by dividing the sum of the ratings for a specific task on a particular scale by the number of raters who rated that task. The mean rating for the Type of Training Scale was computed for a three point scale in which a value of "3" was assigned to a rating corresponding to "Resident School Training", a value of "2" was assigned to those tasks where the rater evaluated the training method most appropriate as being in any of the three following categories: on-the-job training, formal unit training, and non-resident school training. A value of "1" was encoded for the evaluative category corresponding to "no training required". A second type of training scale was a five point scale. The scale values were derived by assigning a value of "5" to the category corresponding to "Resident School Training", a value of "4" to Formal Unit Training", a value of "3" to "Non-Resident School Training", a value of "2" to "On-the-job Training", and a value of "1" to "No Training Required".

A third type of training scale value was generated by assigning a value of "1" to those tasks where the majority of the raters felt that "Resident School Training" or "Formal Unit Training" was most appropriate while a value of "0" was assigned to those tasks that did not meet this criterion.

Another variable included in the analyses reflected the percent of members of the MOS who perform each of the tasks. These data were acquired from previously collected information that was processed in the Comprehensive Occupational Data Analysis Program (CODAP).

Three other variables were generated in accordance with the procedure outlined by Mead (1975). These variables were the squared Task

Learning Difficulty scale values, the squared Task Delay Tolerance scale values and the squared values of the Percent of Members Performing. Generation of these new variables appeared desirable because of the possibility that a quadratic relationship existed between these variables and the Type of Training Scale.

The correlation matrix resulting from the ten variables is shown in Table 3. Since the subsequent analyses of regression indicated that the greatest utility in terms of predicting training priorities resulted from the prediction of training priorities (i.e., TOT) from the four of the predictor variables (i.e. TLD, CIP, TDT, and PCT) the correlation matrix for these five variables is presented in Table 4. The results of

Table 3

CORRELATION AMONG THE THREE TYPES OF TRAINING SCALES AND THE SEVEN PREDICTOR VARIABLES

Variable*	Variable*									
	TOT 3	TOT 5	TOT 1	TLD	CIP	TDT	PCT	TLD PCT	TDT SQ	PCT SQ
TOT 3	1.00	.97	.62	.87	.51	- .11	- .32	.13	- .12	- .27
TOT 5		1.00	.67	.88	.52	- .12	- .38	.14	- .12	- .34
TOT 1			1.00	.50	.30	- .01	- .22	.18	- .01	- .21
TLD				1.00	.56	- .11	- .39	.15	- .13	- .33
CIP					1.00	- .59	- .16	.02	- .61	- .19
TDT						1.00	- .17	.18	.99	.09
PCT							1.00	- .35	- .16	.94
TLD PCT								1.00	.17	- .20
TDT SQ									1.00	- .08
PCT SQ										1.00

*TOT 3 = Three-step Type of Training Scale
TOT 5 = Five-step Type of Training Scale
TOT 1 = Two-point Type of Training Scale
TLD = Task Learning Difficulty
CIP = Consequences of Inadequate Performance
TDT = Task Delay Tolerance
PCT = Percent of Members Performing
TLD PCT = Task Delay Tolerance value if PCT < 5 percent of PCT; otherwise 0
TDT SQ = Task Delay Tolerance Square
PCT SQ = Percent of Members Performing Squared

Table 4

CORRELATIONS AMONG THE THREE-STEP TYPE OF TRAINING SCALE AND THE FOUR PREDICTOR VARIABLES

Variable*	Variable*				
	TOT 3	TLD	CIP	TDT	PCT
TOT 3	1.00	.87	.51	- .11	- .32
TLD		1.00	.56	- .11	- .39
CIP			1.00	.59	- .17
PCT					1.00

*TOT 3 = Type of Training (3 step scale)
TLD = Task Learning Difficulty
CIP = Consequences of Inadequate Performance
TDT = Task Delay Tolerance
PCT = Percent of Incumbents Performing

the six analyses of regression and the corresponding R^2 values shown in Table 5 indicate that the most effective regression equation results from the prediction of training priorities based on a three point scale from ratings on the mean values of four of the scales (i.e. TLD, CIP, TDT, and PCT).

Six analyses of regression were performed using ratings on the Type of Training Scale as the criterion. In the first two analyses, as stated previously, the ratings were trichotomized. In the second two analyses a five point scale was used, while in the third two analyses a dichotomized scale was used as the criterion variable. In each pair of analyses the first analysis employed four predictor variables: three of these variables were the average ratings of each task on the TLD scale, CIP scale, the TDT scale, and the fourth variable was the percent of incumbents performing (PCT) each task in the MOS. The second analysis in each pair of analyses employed three generated variables. One of these generated variables was the value of the TLD value if less than five percent of the members performed the task; otherwise, a value of "0" was assigned to the value of the variable for that task. The two other variables were the squared TDT ratings and the squared PCT values for each task.

The beta weights obtained in each of the six analyses are shown in Table 5 accompanied by the corresponding multiple correlation coefficients and the squared multiple correlation coefficients.

Table 5

BETA WEIGHTS OF THE DIFFERENT VARIABLES IN THE SIX REGRESSION ANALYSIS FOR PREDICTING TRAINING PRIORITIES

Variable*	Type of Training 3 pt scale		Type of Training 5 pt scale		Type of Training 2 pt scale	
TLD	.85644	.85186	.83094	.83393	.44552	.46658
CIP	.04597	.05615	.05028	.07647	.11853	.11459
TDT	.01765	-.35526	-.00236	1.25900	.10873	-.48737
PCT	.01909	-.07334	-.04890	-.08289	-.00728	.35399
TLD PCT		-.00576		-.00380		.14626
TDT SQ		.37276		1.27285		.60386
PCT SQ		.09161		.02835		-.32794
R	.87330	.87393	.88019	.88392	.51012	.53106
R^2	.76266	.76375	.77473	.78131	.26022	.28203

*TLD = Task Learning Difficulty
CIP = Consequences of Inadequate Performance
TDT = Task Delay Tolerance
PCT = Percent of Members Performing
TLD PCT = TLD of $\leq$ 5%; otherwise TLD
TDT SQ = TDT squared
PCT SQ = Percent of Members Performing squared

The highest multiple correlation coefficient was obtained using the five point type of training scale as the criterion and using all seven predictor variables; however, there was only a negligible increase in this instance when compared with using the four original variables. Also, it should be noted that a negligible difference existed between the multiple correlation of the three-point type of training scale and the five-point type of training scale and that the difference in predicting the former from the four or the seven predictor variables was

not of substantial importance. A substantial difference did exist when attempting to predict the dichotomized type of training rating from either the four or from the seven prediction variables. Generally, then it appears that the most parsimonious policy capturing can be achieved by predicting the trichotomized type of training scale from the four original variables.

Generally, the results of the interviews indicated that the task listing and instructions for the use of the scales were adequate. However, there appeared to be some problem with the scale reflecting Task Delay Tolerance. As noted previously this scale was reversed in terms of importance from the direction of the other scales. The problem of being assigned to duty in an MOS other than their primary MOS was of concern to some of the non-commissioned officers particularly in view of the fact that they were required to take their proficiency test in their primary MOS. Many NCO's were trained in this MOS by on-the-job training and some of these were partial to the OJT training strategy. Some non-commissioned officers felt that school acquired training was not as utilized as it should be because of locally developed procedures or because supervisors were not fully aware of the school curriculum. Another reason for not fully utilizing school acquired information was that supervisors were not aware of the current curriculum in the school.

DISCUSSION

The results of this research indicated that the four factor model of determining training priorities is applicable to MOS 76V Equipment Storage Specialist. The single most useful variable in the policy capturing equation was the task learning difficulty scale. In other words, tasks that are rated as being difficult to learn should be taught in a more formalized environment. Again, from the standpoint of assessing training priorities the most parsimonious approach appears to result from segmenting the type of training variable into a three point scale that corresponds to resident school instruction, other types of training and no training required. The four basic variables (i.e. Task Learning Difficulty, Consequences of Inadequate Performance, Task Delay Tolerance, and Percent of Members Performing) are highly predictive of the type of training when classified in this fashion and there is no apparent need for the three generated variables since they do not add sufficiently to the prediction equation.

The relatively low predictive effectiveness of the Task Delay Tolerance Scale in predicting type of training may be attributed to the relatively low reliability of that scale. As noted previously, this was possibly a function of this scale having step values that were reversed in comparison with the other three scales. Further research will be conducted to determine if this scale has greater utility when the scale step values are the same as on the other three scales.

Other analyses of the data obtained in this research will aim at isolating the differences in the ratings provided by supervisors and instructors as well as those differences that may be attributable to the order of presentation of the scales. Also, it would be interesting to explore the increase in prediction that may result from different scaling strategies. It is recommended that the utility of the four factor model be examined in other MOS before the system be made fully operational in the Army.

SUMMARY

The purpose of this research was to explore the feasibility of a strategy for assigning training priorities in an Army supply MOS. Data were collected from supervisors and instructors for each of the 183 tasks in the MOS on four rating scales: task learning difficulty, consequences of inadequate performance, need for immediate performance, and the most appropriate type of training. Additional data were obtained for the tasks that reflected the percentage of members performing each task in the MOS. Data were analyzed in terms of the reliabilities of the four scales and regression analyses were performed to establish the degree to which training priorities could be predicted from four variables (i.e. task learning difficulty, consequences of inadequate performance, need for immediate performance, and percentage of members performing). A multiple correlation coefficient of .87 was obtained between the type of training scale and these four variables in the most parsimonious solution. Additional analyses of these data will be performed to isolate certain effects which may lead to increased utility of the model. Further research will involve exploring the value of the model in other Army MOS and examining its feasibility for operational use on an Army-wide basis.

REFERENCE

Mead, D. F. Determining training priorities for job tasks. Paper presented at the 17th Annual Conference of the Military Testing Association, Indianapolis, Indiana, 16-19 September 1975.

Development of CAI Performance Measures: TACFIRE Tactical Data System

W. G. Hoyt, A. K. Butler, P. W. Leung
System Development Corporation

and

J. E. Germas and J. T. Larson
U.S. Army Research Institute

Army learning is becoming more decentralized, placing greater emphasis on unit and individual training programs and capitalizing on the inherent computer capabilities of Army Tactical Data Systems.

In May 1975 the System Development Corporation was tasked by the U.S. Army Research Institute to develop embedded individualized training programs for users of the TACFIRE Tactical Data System, a computerized field artillery command and control system. The computer assisted instruction (CAI) language used was PLANIT (Programming Language for Interactive Teaching).

The five major phases of this project are shown in Figure 1.

Phase I	- Analyze System and Training Requirements
Phase II	- Perform Job/Task and Training Analysis
Phase III	- Develop Courseware
Phase IV	- Install Courseware
Phase V	- Develop Field Evaluation Plan

Figure 1. Major Project Phases

Forty hours of TACFIRE CAI courseware and performance tests were developed under this project, and the training program is currently operational on the TACFIRE system. The program was developed to train the Battalion Fire Direction Officer and Fire Direction Sergeant in TACFIRE Fire Direction Center (FDC) operations. The various FDC's are

the focal points in the TACFIRE system, as shown in Figure 2. The Artillery Control Console, Figure 3, is used to conduct FDC operations and also to administer the computer assisted instruction (CAI). This paper reports on the development of CAI performance measures of job proficiency for these TACFIRE Tactical Data System users.

Each of the phases in Figure 1 has an impact on the development of measures of job proficiency. The first four phases are concerned with the development of the concepts, principles, procedures, tasks and products that are part of the measurement development process. The fifth phase specifies how performance measures will be administered. These concepts, principles, and procedures are relevant to the development of any job based performance measures, whether they are administered in a CAI mode or in another mode. The contribution of the first three phases is covered in the paragraphs that follow.

Phase I - Analyze System and Training Requirements

The process of developing performance measures starts with the analysis of system and training requirements. The analysis provides the guidelines, principles, and framework for measuring performance.

For example, the analysis of the TACFIRE computerized system shows that it provides rapid and accurate transmission and utilization of field artillery data. Computerized control, as employed in TACFIRE, involves a great deal of specificity: precise inputs, and the filling in of required fields and subfields that allows little margin for error. The formats used are many and varied and the data entries required for each format are extensive and precise, as shown in Figure 4. There are over one hundred of these formats, as listed in Figure 5. Initially, it appears that a retentive memory and an extensive memory span are required of TACFIRE users and should be tested for. However, the analysis shows that this is not the case. TACFIRE fire mission and data bases--guns, ammo, target, meteorological, etc.--are input from many different units and agencies (see Figure 2), either directly into the computer or by oral or written communications. Consequently, much of the TACFIRE operation involves the processing--rahter than the generation--of data, i.e., recognition rather than memory--recall. The operational requirement becomes (1) processing of completed message formats or (2) the association of input data with a specific format, calling up the format, making the entries specified in the format, and transmitting the data to the units requiring them.

Thus, the analysis provides guidelines for the development of job performance measures and specifies what should be included in the measurement situation. Much of the performance measurement problem can

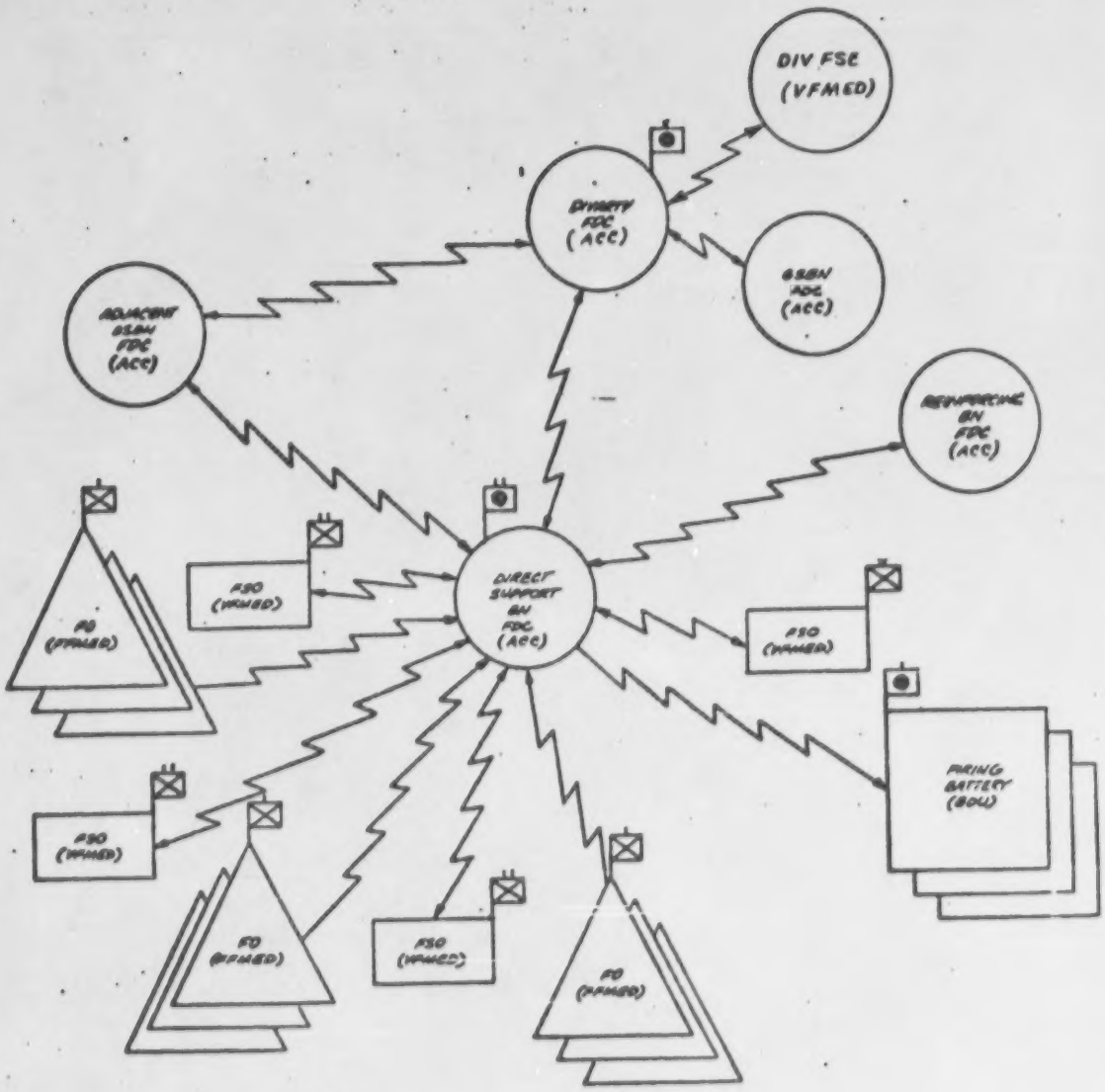


Figure 2. Location of TACFIRE Operations Personnel and Devices

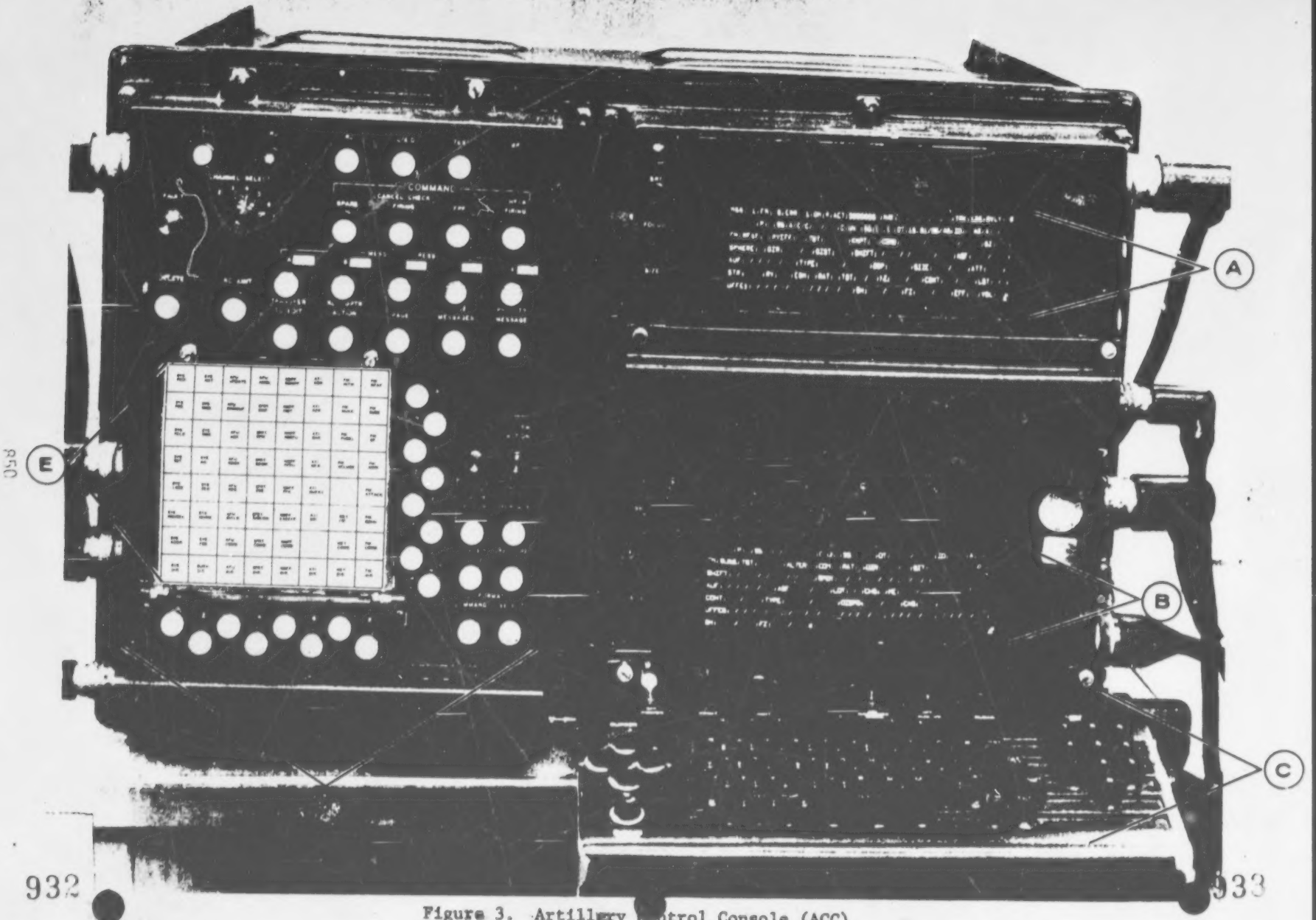


Figure 3. Artillery Control Console (ACC)

be alleviated by maintaining the job orientation in measuring performance, i.e., measure performance of the individual in the context in which he will do the job.

```
      ;P: ;SB: / / /  /   ;C:   ;SG:  ,  ;DT:  ,  /  /  ;ID:    ;A: ;
AFU;BAMOUP;FU: / /C/1 /4Ø ;AMOR: ;AMOE: ;AMOH:X;PLAN:       ;STDODD:H/O;
PROJA:HEA1/H/ 33.Ø/6Ø6 ,HEC1/F/ 33.Ø/6Ø6 ,SMA1/S/ 33.Ø/133 ;
PROJB:SMB1/M/ 33.Ø/228 ,SMC1/E/ 33.Ø/6   ,SMD1/K/ 33.Ø/12  ;
PLOT:M67 /H/6Ø6 ,M67 /Ø/1Ø44,M67 /F/14Ø1:DTG:  /  /  ;
FZES:PDA /282 ,PDC /6Ø  ,TIA /132 ,TIB /678 ,TIJ /258 ,VTC /36Ø ;BKUP: ;
MYIELD:    /      /  ,   /      /  ,   /      /  ,   /      /
```

Figure 4. Example of a TACFIRE Message Format - AFU; BAMOUP

Phase II - Job/Task and Training Analysis

The steps in Phase II are shown in Figure 6. The TAIS (Figure 7) specify the tasks to be trained and on which performance will be measured. The TAIS records the results of the training analysis and provides behavioral task information leading to the objectives, test items, and course materials. Note that the job orientation is maintained, including the functional relationship between message formats. The criterion and enabling objectives are shown in Figure 8. Criterion objectives specify the type of behavior required for the task and lead directly to the test item. Note the audit trail is maintained by the TAIS number, 3004, in the upper left corner. The test items shown in Figure 9 are developed directly from the criterion and enabling objectives. The guidelines used in developing test items are shown in Figure 10. At this point "what is to be measured" and "how" have been defined. The TAIS, Criterion and Enabling Objectives, and Test Items were reviewed and approved by TACFIRE subject matter experts at the U.S. Army Field Artillery School prior to Phase III, courseware development.

Phase III - Develop Courseware

The design of each course module (functional area) is shown in Figure 11. Performance measures include Form A (Pre-test), Form B (Post-test) of the Module tests (middle section), the criterion test element within each lesson (bottom section), and individual items within the instructional element. The evaluation element of each lesson evaluates the test scores (done automatically), provides the results and branches the student to the next lesson or to a remedial path depending

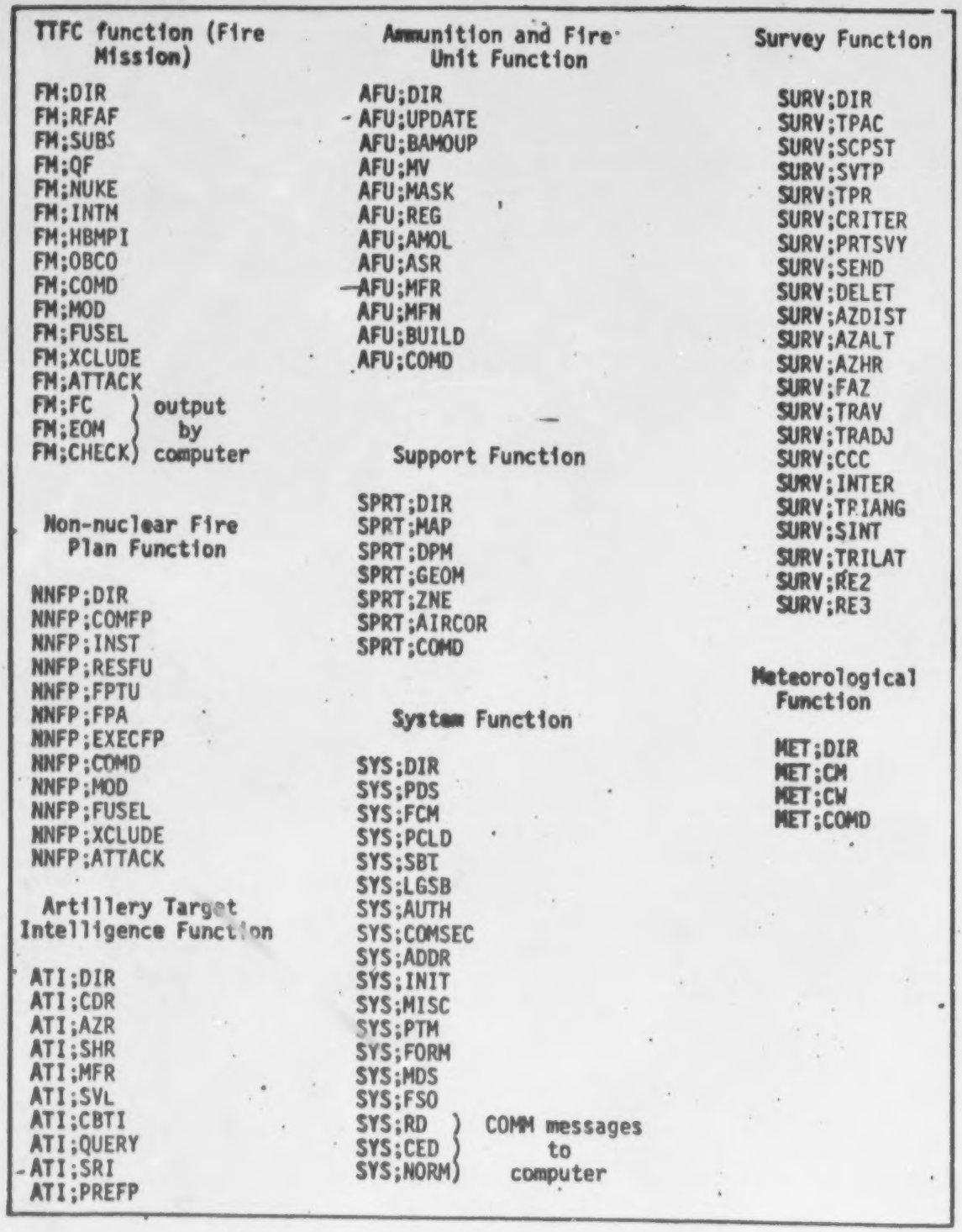


Figure 5. TACFIRE Message Formats

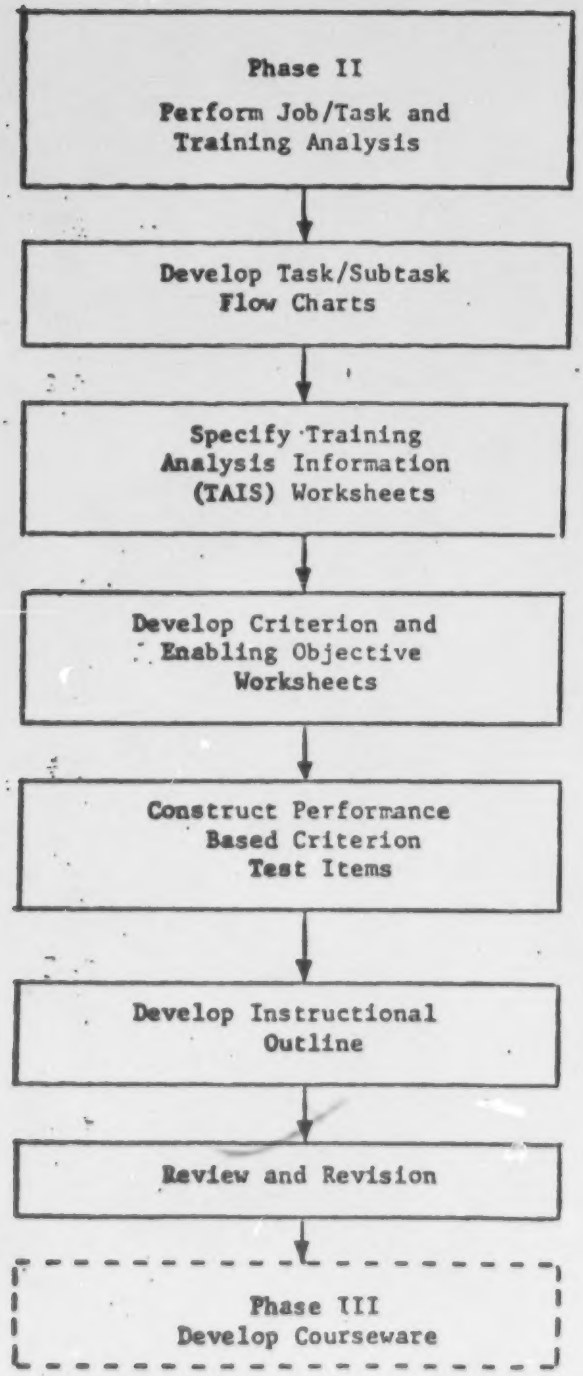


Figure 6. Phase II - Job/Task and Training Analysis-Developmental Steps

TAIS No. 3004

MODULE AFU

UNIT BAMOUP

TRAINING ANALYSIS INFORMATION SHEET

1. TASK IDENTIFICATION: 4.0

2. TASK: Update the ammunition inventory for an active FU to reflect ammunition received and verify data entries.

3. CONDITIONS: Given requirements to update current ammunition status of a FU to reflect ammunition received, select correct message format and fill in appropriate entries. Given the requirement to print and interpret AFU 2204 FU AMMO SUMMARY output message, select correct message format to print output message and interpret contents. Given different formatted test items concerning the updating of the ammunition status for a FU and AFU 2204 FU AMMO SUMMARY output message, provide correct response.

4. STANDARD: No errors.

5. TASK ANALYSIS:

TASK ELEMENTS	PREREQUISITE KNOWLEDGE OR SKILL REQUIREMENTS	SUPPLEMENTAL TRAINING MATERIAL	REFERENCES
4.1 Select and display AFU;BAMOUP massage.	4.1 Know operation of ACC component parts.	1. Picture/ drawing of ACC.	DTM 11-7440-240-10
4.2 Identify entries for ammunition data.	4.2 Know operation of ACC component parts.	2. Entry data and AFU; BAMOUP format.	Chapter 4 Pages 4-159 through 4-176D.
4.3 Identify results of computer action.	4.3 None.	3. Picture of AFU 2204 FU AMMO SUMMARY output message.	Chapter 6 Pages 6-1 through 6-6; 6-21 through 6-26; 6-75 through 6-103.
4.4 Print AFU 2204 FU AMMO SUMMARY output message.	4.4 Know how to select and display AFU user command message.	4. Additional material to be developed as required.	
4.5 Interpret AFU 2204 FU AMMO SUMMARY contents.	4.5 Able to decode mnemonics.		

Figure 7. TAIS for a TACFIRE Functional Area Topic: Ammunition and Fire Unit (AFU) Function

TAIS No. 3004

MODULE AFU

UNIT BAMOUP

CRITERION AND ENABLING OBJECTIVES

TASK IDENTIFICATION: 4.0

TASK ELEMENTS: 4.1 - 4.5

CRITERION OBJECTIVE(S)	ENABLING OBJECTIVE(S)
4.1 Given a picture/drawing of the ACC switch panel assembly, identify the switch actions that can be used to select and display the AFU;BAMOUP message. The switch matrix is referenced by letters for rows and numbers for columns. The student is able to match the correct letter/number combination to select the required message format. The correct steps are: a. Depress switches B and 3. b. Activate FORMAT COMMAND switch.	4.1.1 Pick from a list the purpose of the AFU;BAMOUP message as being: MAINTAIN AMMUNITION DATA FOR A FU.
	4.2.1 State ON THE CED as being where the AFU;BAMOUP will display after being selected.
4.2 Given information to update the ammunition status of a FU to reflect ammunition received, the student will identify the data to simulate the completion of the AFU;BAMOUP input message. Data entries will include: • Fire Unit • Ammunition received • Ammunition characteristics • Powder characteristics (Data to be specified)	4.2.2 Match the following mnemonics with their definition. a. FU - Fire Unit b. AMOR - Ammunition received c. PROJA - Ammunition characteris- PROJB tics d. PLOT - Powder characteristics Note: Explanation of additional mnemonics will be included within the instructional material for student review. 4.2.3 Pick from a list the entry that is always required in the AFU;BAMOUP message as being: FU.

Figure 8. Criterion and Enabling Objectives Worksheet: AFU Function

TAIS No. 3004

MODULE AFU

UNIT BAMOUP

TEST ITEMS

TASK IDENTIFICATION: 4.0

TASK ELEMENTS: 4.1 - 4.5

CRITERION ITEM(S)	ENABLING ITEM(S)
4.1 Refer to Figure ___. Assume one of the fire units in your Bn has received additional supplies of ammunition which must be added to their ammunition inventory. As a first step you need to select the message format so that you can enter this information and update the ammunition status for the FU. From the list of steps given below, first select the procedural steps required and then place them in the correct order. a. Activate FORMAT SELECT switch. b. Activate FORMAT COMMAND switch. c. Activate REPLACE switch. d. Depress switches G and 3. e. Depress switches B and 3. (e, b)	4.1.1 The AFU;BAMOUP message is used to: a. Maintain data on backup units. b. Assist Bn in maintaining Battery availability files. *c. Maintain ammunition data for a FU d. Set amount of ammunition that can be expended by each Battery. 4.2.1 After being selected by the appropriate switch actions, the AFU; BAMOUP message format will appear on the (RD/CED)?
4.2 One of your FUs has received a supply of ammunition to replace previously expanded ammunition. You have already selected the AFU; BAMOUP format message (Figure C) and must now enter the information to update the ammunition inventory for the FU. Answer the following questions concerning the entry of data into the AFU;BAMOUP message format . (Sample data and questions) Data to be included in Figure D. • Btry B, 1st Bn, 41st Regiment.	4.2.2 From the following list, match each mnemonic with its definition and function. a. Powder characteristics b. Fire Unit c. Ammunition received d. Ammunition characteristics FU (b) PROJB (d) PLOT (a) PROJA (d) AMOR (c)

Figure 9. Test Item Worksheet: AFU Function

- Test items must be performance-based and job-oriented and require the student to demonstrate skills and knowledge directly related to the criterion objectives.
- Each test item must elicit measurable behavior.
- The structure of the test item must be positively oriented.
- Multiple-choice items must have at least four alternatives.
- The test item must be amenable to CRT presentation or CRT presentation plus a simple off-line exhibit.

Figure 10. Guidelines for Developing Test Items

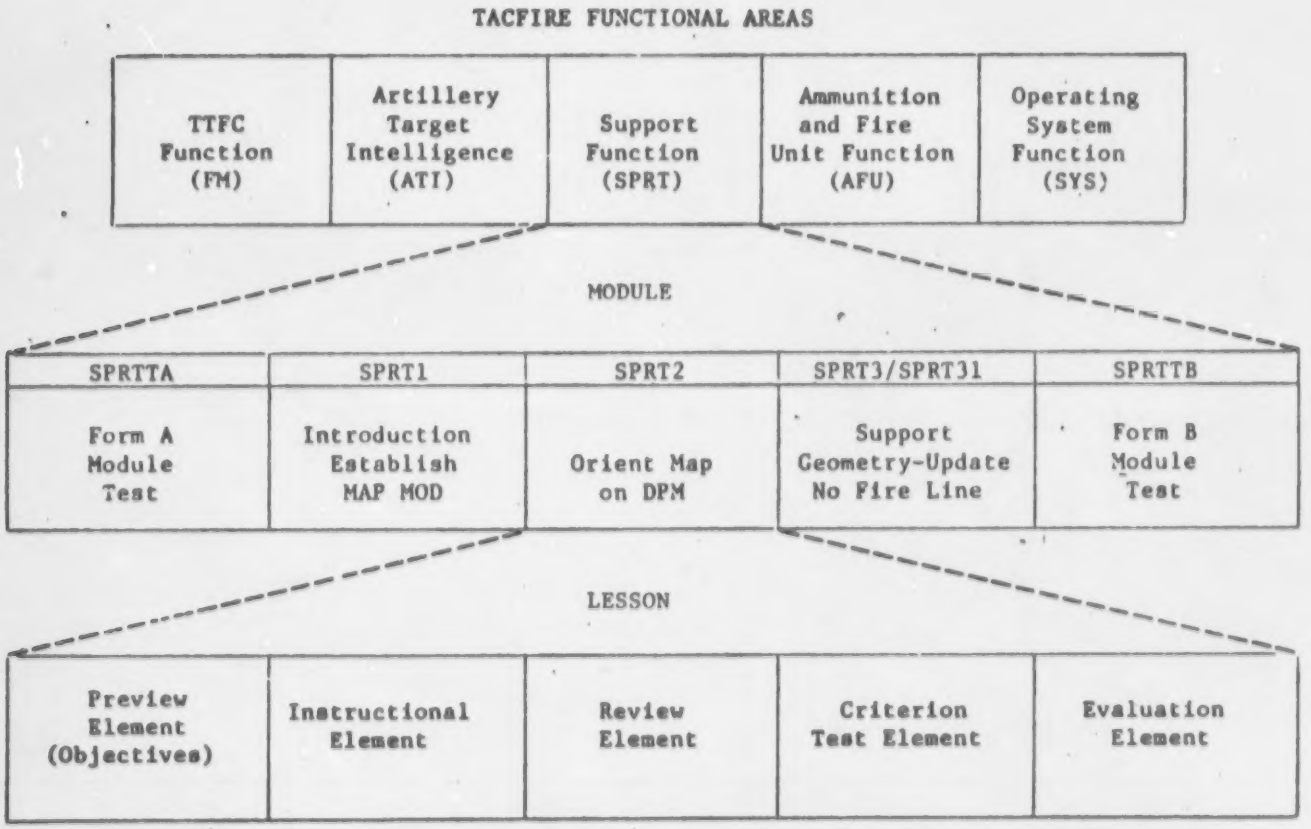


Figure 11. Organization of TACFIRE AI Courseware

858

941

942

- Provide all the relevant cues and stimuli that are part of the tactical operation.
- Use context examples of TACFIRE message formats, displays and messages that are relevant to the TACFIRE operational environment.
- Maintain functional context by testing TACFIRE operations and message formats in the order they are encountered on the job.
- Provide clear instructions as to the situation and what the task is.

Figure 12. Ground Rules for Developing TACFIRE Performance Measures

a. FRAME 3.00

IN CONDUCTING TACFIRE OPERATIONS, YOU WILL BE SITTING AT THE ACC, OBSERVING THE DISPLAYS ON THE RD AND CED, AND USING THE SPA AND KEYBOARD TO CARRY OUT YOUR ACTIONS. THIS JOB SITUATION PROVIDES MANY CUES AS TO WHAT ACTIONS ARE TAKEN AND HOW THEY ARE TAKEN. FOR EXAMPLE, YOU DO NOT NEED TO MEMORIZE THE NAME OF A SWITCH. ALL YOU NEED TO DO IS PICK IT OUT ON THE SPA. (TYPE GO)

b. FRAME 4.00

USE THIS JOB SITUATION IN TAKING THE TEST. PICTURES OF THE ACC (FIGURE 1) AND THE SPA (FIGURE 2) ARE INCLUDED IN THE TEST EXHIBITS. REFER TO THEM WHENEVER YOU LIKE AND USE THEM TO "VISUALIZE" THE STEPS REQUIRED IN CARRYING OUT THE ACTIONS REQUIRED. NOW FOR THE TEST QUESTIONS.

Figure 13. CAI Frames Referencing the Tactical Environment and Cues

a. FRAME 19.0

YOUR BN HAS RECEIVED A SUPPLY OF AMMUNITION AND THIS INFORMATION NEEDS TO BE ENTERED INTO THE TACFIRE COMPUTER. YOU HAVE ALREADY SELECTED AND DISPLAYED AN EMPTY AFU;BAMOUP MESSAGE FORMAT (FIGURE 7) AND NOW NEED TO ENTER THE DATA (FIGURE 6) TO UPDATE THE AMMO STATUS FOR THE FIRE UNIT.

(TYPE GO)

b. FRAME 20.0

REFER TO FIGURES 6 AND 7. WHICH OF THE FOLLOWING IS THE CORRECT ENTRY TO SPECIFY THE FIRE UNIT?

A. FU: /B/1/41/ ;
B. FU:B/ / / /141;
C. FU: / /B/1 /41 ;
D. FU: / /1/41/B ;

(ENTER A LETTER)

c. FRAME 21.0

REFER TO FIGURES 6 AND 7. WHICH OF THE FOLLOWING IS THE CORRECT ENTRY FOR ENTERING THE AMMUNITION CHARACTERISTICS RECEIVED?

A. PROJA: 30/X/ILA.1/ 33,
B. PROJA:ILA1/I/ 30. /33.0,
C. PROJA: /X/ 33.0/ILA1,
D. PROJA:ILA1/I/ 33.0/ 30,

(ENTER A LETTER)

Figure 14. CAI Test Frames Presented on CRT.

- FIRE UNIT: B BTRY, 1/41ST
- ACCOUNTING PROCEDURE: AMMUNITION RECEIVED
- AMMUNITION CHARACTERISTICS

  CATEGORY: 1LA1
  LOT DESIGNATOR: I
  WEIGHT: 33.Ø
  QUANTITY: 3Ø

- POWDER CHARACTERISTICS

  MODEL: M67
  LOT DESIGNATOR: X
  QUANTITY: 1ØØ

- FUZE CHARACTERISTICS

  CATEGORY: TIB
  QUANTITY: 3Ø

Figure 6. Ammunition Message

```
     ;P: ;SB: / / /  /   ;C:   ;SG:  ,  ;DT:  ,  /  /  ;ID:    ;A: ;
AFU;BAMOUP;FU: / / /  /   ;AMOR: ;AMOE: ;AMOH: ;PLAN:      ;STDODD: / ;
PROJA:    / /   . /    ,   / /   . /    ,   / /   . /    ;
PROJB:    / /   . /    ,   / /   . /    ,   / /   . /    ;
PLOT:    / /    ,   / /    ,   / /    ;DTG:  /  /  ;
FZES:     /   ,   /   ,   /   ,   /   ,   /   ,   /   ,   /   ;BKUP: ;
MYIELD:    /      / .   /      / .   /      / .   /      /
```

Figure 7. Emoty AFU; BAMOUP Displayed on CED

Figure 15. Ammunition Message and Empty AFU;BAMOUP Message Format--Off-Line Course Exhibit

upon his performance. In developing the lesson and module performance measures certain ground rules (shown in Figure 12) were observed. These ground rules help establish a close, one-to-one relationship between the job situation and the performance measurement situation.

The development of the CAI sequence used to measure operational performance adheres to the Task Elements in the TAIS, criterion and enabling objectives, and test items as shown in Figures 7, 8, and 9.

At the start of each test, the two frames shown in Figure 13 are presented on the CRT. These alert the student to the cues provided by the operational situation and reference the pictures of the artillery control console provided in the off-line course exhibits.

Just prior to this sequence, in Task Elements 4.1, the performance measured was the ability to associate the incoming ammunition message with the AFU;BAMOUP message format and to select and display the format. In the CAI sequence for Task Element 4.2 shown in Figure 14, the performance measured is correct completion of the AFU;BAMOUP message format. The tactical message and the ACC format display are provided to the student to answer the questions posed by the CAI Text. This information is shown in Figure 15 as it appears in the off-line course exhibits. This combination of CAI Text and exhibits provides an accurate operational environment and a tactical situation in which the student's job performance can be measured. It should be noted that the audit trail is maintained in that the header line of each CAI test frame (which the student does not see) includes the TAIS identifier, i.e., AFU 4.2

In conclusion, an examination of Task Elements 4.1 and 4.2 in the TAIS 3004 (Figure 7) indicates that the performance that must be measured in order to assess whether Task Elements 4.1 and 4.2 have been properly addressed is that shown in Figure 16.

The ability to:

- Associate an incoming message with the required message format
- Select and display the message format
- Correctly fill in the format
- Process the format by taking computer action and transmitting to the units requiring them

Figure 16. Performance Measured

These performance measures are those suggested in the Phase I analysis and delineated in Phases II and III. The development cycle is thus completed. Systematic application of the notion of CAI (or non CAI) performance measures of job performance within the operational context has produced a technique to permit development of additional valid measures of job performance.

REFERENCES

1. Hoyt, W. G. Utilization of Tactical Computers for Training: Analysis of System Training Requirements. TM-5544/000/00, System Development Corporation, Santa Monica, CA., 20 June 1975.

2. Butler, A. K., Hoyt, W. G., Leung, P. W. Utilization of Tactical Computers for Training: Job/Task and Training Analysis. TM-5544/001/00, System Development Corporation, Santa Monica, CA., 20 August 1975.

3. Hoyt, W. G., Butler, A. K., Leung, P. W. Utilization of Tactical Computers for Training: Field Evaluation Plan. TM-5544/002/00, System Development Corporation, Santa Monica, CA., 5 December 1975.

4. Butler, A. K., Hoyt, W. G., Leung, P. W. Utilization of Tactical Computers for Training: Job/Task and Training Analysis, Ammunition and Fire Unit (AFU) Module. TM-5544/001/01, System Development Corporation, Santa Monica, CA., 1 March 1976.

5. Hoyt, W. G., Butler, A. K., Leung, P. W. Utilization of Tactical Computers for Training: Final Report. TM-5544/003/00, System Development Corporation, Santa Monica, CA., 5 May 1976.

6. Germas, J. E., Johnson, E. M., Baker, J. D. Embedded Training: Using a Tactical Computer System to Train the System Operator. Paper presented at the Sixth Congress of the International Ergonomics Association, College Park, Maryland, 11-16 July 1976.

TESTING OF SOUTHEAST ASIA REFUGEES TO DETERMINE MILITARY ENLISTMENT ELIGIBILITY

BY

Major Wayne S. Sellman
Air Force Military Personnel Center

AND

Captain Melvin B. Gambrell*
Armed Forces Vocational Testing Group

During the summer of 1975, considerable concern was expressed by the Inter-agency Task Force on Southeast Asian Refugee Resettlement because the rate of resettlement was far below expectations. One solution discussed was that the Department of Defense should sponsor legislation for refugees waiving the one-year statutory residence requirement and permitting Service enlistment eligibility for those refugees meeting minimum enlistment standards. In short, qualified refugees not resettled into the civilian sector would be allowed to enlist in the military. However, at that time, no data were available to indicate just how many refugees that might be. Accordingly, the Office of the Assistant Secretary of Defense (Manpower and Reserve Affairs) requested the Air Force to obtain information concerning the English language proficiency levels and aptitude profiles of refugees at Fort Indiantown Gap, Pennsylvania.

APPROACH

A team from the Air Force Military Personnel Center and the Armed Forces Vocational Testing Group visited the refugee center during 8-12 September 1975. A representative sample, drawn from the 18 to 35 age group, was administered the English Comprehension Level Screening Test (ECLST). The test, which measures listening and reading comprehension skills, was developed by the Defense Language Institute and validated through its use with thousands of foreign military personnel. To preclude any possible chance of test compromise, four parallel versions of the ECLST were used. (See DLI Pamphlet 350-3, English Language Training Program, Programming of Training and Administration of ECL Screening Tests, for a description of the ECLST.)

*Now serving as USAF Exchange Officer with the Royal Australian Air Force, Headquarters Support Command, Melbourne, Victoria, Australia.

Those refugees who demonstrated a sufficient competence with English were administered the Armed Services Vocational Aptitude Battery (ASVAB) (Bower & Lewis, 1975; Sellman, 1975; Valentine, 1975; Wilfong, 1975; and Jensen & Valentine, 1976) to determine eligibility for enlistment into the various Armed Services. A cut-off score of 55 on the ECLST, which represents a working level proficiency in reading, writing, speaking, and comprehension, was established as a requirement for ASVAB testing. This cut-off is five points lower than the minimum score required by the Army for entry of foreign military students into training programs and fifteen points lower than standards for the other Services. It should be noted that this minimum is for the relatively unsophisticated courses. The more complex the skills to be learned, the higher the ECLST score required. The "passing" score was reduced in order to insure that no one who might possess even a minimum language proficiency was denied ASVAB testing.

The initial sample was administered a demographic/biographic questionnaire designed to verify (through self-report) critical variables including age, sex, marital status, education level, prior military service, prior English language training, prior civilian employment, and future occupational desires. Instructions for completing the questionnaire were translated into Vietnamese and Cambodian and were read to the examinees by interpreters.

The testing sessions were conducted under standardized conditions. Two testing rooms of identical configuration were used. They were properly lighted and ventilated, and of sufficient size to adequately test 40 individuals per session. Representatives from the Defense Language Institute administered the ECLST while the ASVAB was given by representatives of the Armed Forces Vocational Testing Group.

The testing team was careful not to raise the expectations of the refugees or create the impression that the Department of Defense was interested in becoming a sponsor. Prior to testing, an article, which briefly described the project as an exercise to determine the English language proficiency and vocational aptitudes of refugees awaiting sponsorship, was published in the camp newspaper. No mention was made to the refugees of any military orientation to the project. As an additional safeguard, test booklet covers were removed before testing.

FINDINGS AND DISCUSSION

The original sample to be tested was selected based on percentages of refugees at Fort Indiantown Gap by sex, age and education level during the week 25-29 August 1975. Refugees were identified by name and asked to participate in the project. However, between the time the refugees were identified and the time testing occurred, approximately 2,200 of the Indiantown Gap refugees were resettled, including many earmarked for the project. Statistical analyses (chi square) were computed to determine if the resettled refugees differed significantly from those originally selected for the project. Generally, those refugees resettled possessed higher levels of education than those remaining in the camp. Further, to obtain a sufficiently large sample which would allow inferences about English proficiency and aptitude profiles, additional refugees were identified and asked to participate. Again chi square analyses were computed to determine if the sample actually tested met all the sex, age and education characteristics of the original one. Results indicated that the sample used in the project at the time of testing was highly representative of the camp's population in age groups 18-35.

Three hundred twenty-eight refugees participated in the testing project. Demographic/biographic information which describes the sample is presented in Tables 1-9 in the Appendix. It should be noted that not all entries in the tables sum to the sample size 328. While an attempt was made by the testing team and interpreters to ensure complete and accurate responses to items within the questionnaire, due to language difficulties and time constraints not all refugees completed each item. There was, however, adequate response to allow the identification and interpretation of trends.

Of those refugees tested, approximately 11% achieved a passing score on the ECLST. Tables 10 and 11 display the percent male and female by age and education level who qualified. Inspection of the two tables reveals that only a small percentage of those tested demonstrated even a minimum proficiency with the English language.

To investigate the relationship between information collected in the demographic/biographic questionnaire and ECLST scores, a series of statistical analyses were computed. Generally, regression analyses failed to predict English proficiency scores from questionnaire data. As might be expected, the level of education and

the length of English language training were the only two background variables which demonstrated any correlation with refugee performance on the ECLST. Table 12 presents the correlation coefficients of each demographic/biographic variable with English language proficiency scores.

Approximately ten percent of the 71% of the sample who indicated some prior military service passed the ECLST (Male prior military service = 10.38%; female prior military service = 00%). Table 13 shows the percent male and female by age and prior military service who obtained a passing score on the English proficiency test. The correlation between prior military service and English proficiency was negligible (r = .04). In responding to the questionnaire, refugees were asked to indicate what kind of job they were interested in upon resettlement. Only four percent of the overall sample reported an interest in the military service (see Table 8). Within that group only one individual passed the ECLST. In percentage terms, .3049% of the overall sample tested indicated interest in the military service and possessed minimum English language skills for entry into a military training program.

The Defense Language Institute has determined the length of English language training required for a foreign military student to achieve an English comprehension level suitable for entry into the various military training programs (DLI Pamphlet 350-3, English Language Training Program, Programming of Training and Administration of ECL Screening Tests). Table 14 presents the Defense Language Institute's recommended length of training times. To obtain the minimum ECLST cut-off score for entrance into Army training courses, 50% of the male refugees and 48% of the females would require 23 weeks of English language training as taught at the English Language School of the Defense Language Institute. For that group, 30 weeks of training would be required for entry into training provided by the other Services. Tables 15 and 16 display the percentage distributions of male and female refugees by age who would require extensive English language training before accession into the Armed Forces. Cross-referencing with Table 14 allows determination of the amount of training needed by the various groups. DLI Pamphlet 350-3 provides the language proficiency code key which allows comprehensive interpretation of ECLST scores.

The Armed Services Vocational Aptitude Battery (ASVAB) was administered to those refugees who passed the ECLST (11%). Of those ASVAB tested, four percent of the original sample achieved aptitude scores which would qualify

them for enlistment. Table 17 presents a summary of mental test score criteria for enlistment into the various Armed Services. Table 18 displays the number and percent of refugees from the original sample who qualified by branch of service (USA - 7%; USN - 8%; USAF - 4%; USMC - 5%; USCG - 8%). Review of this table indicates that while individual services vary, only a small percentage achieved sufficient aptitude levels to meet Service accession requirements. Further, Table 19 provides information concerning refugee performance by age, sex, and mental category. Inspection of this table reveals that males achieved higher mental category distribution than did females.

In summary, the greatest handicap to refugee eligibility for the Armed Forces was their English language proficiency. Eighty-nine percent of the refugees tested failed to meet minimum language competency levels. Of the 33 refugees who did possess English competency (completed the ASVAB), 85% achieved enlistment qualification for at least one of the branches, with 33% qualifying for all branches of the military.

CONCLUSIONS

As it turned out, all refugees, awaiting resettlement at the time of this study, were eventually placed into civilian jobs in communities around the country. Thus, the necessity to use the military as a sponsor for the refugees never materialized. However, to the extent that the refugees at Fort Indiantown Gap were representative of refugees at the other camps, the following conclusions were reached.

a. Only a very small number of those refugees remaining in camps would have qualified for enlistment in the Armed Services. Those refugees with more education and greater language facility would have been easier resettled so as time passed the percentages mentally eligible for enlistment would have declined.

b. At the time of the study, there were relatively few refugees interested in the military service as an occupation. However, as chances for resettlement appeared to diminish, the numbers of refugees willing to enter the military might have increased dramatically.

c. With the exception of those individuals who entered semi-skilled jobs, extensive language training was required for most refugees before they could successfully complete any kind of technical/vocational training (civilian or military).

d. Given adequate English language proficiency, a fairly substantial percentage of refugees possessed sufficient aptitude profiles for military eligibility.

REFERENCES

Bower, G. I. and Lewis, J. R. A concurrent validity study relating the Armed Services Vocational Aptitude Battery to success in high school vocational technical courses. AFVTG-TRN-74-5. Randolph AFB, TX: Armed Forces Vocational Testing Group, January 1975.

--------. English language training program: programming of training and administration of ECL screening tests. DLI Pamphlet 350-3. Washington DC: HQ Defense Language Institute, November 1973.

Jensen, H. E. and Valentine, L. D. Validation of ASVAB-2 against civilian vocational - technical high school criteria. AFHRL-TR-76-16. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, March 1976.

Sellman, W. S. (Chairman). Use of a common aptitude test for entry into all military services. Symposium presented to the 17th Annual Conference of the Military Testing Association, Indianapolis, IN, September 1975.

Valentine, L. D. Development of the Armed Services Vocational Aptitude Battery, Forms 5, 6, and 7. Paper presented to the 17th Annual Conference of the Military Testing Association, Indianpolis, IN, September 1975.

Wilfong, H. D. High school use of ASVAB. Paper presented to the 17th Annual Conference of the Military Testing Association, Indianapolis, IN, September 1975.

SELECTED BIBLIOGRAPHY

Burkett, J. R. and Huff, K. H. Survey of Vietnamese air technical training in United States Air Force technical schools. AFHRL-TR-70-17. Lowry AFB CO: Technical Training Division, Air Force Human Resources Laboratory, June 1970.

Guthrie, G. M., Brislin, R., and Sinaiko, H. W. Some aptitudes and abilities of Vietnamese technicians: implications for training. IDA Paper P-659. Arlington VA: Science and Technology Division, Institute for Defense Analysis; November 1970.

Guthrie, G. M., Sinaiko, H. W., and Brislin, R. Nonverbal abilities of Americans and Vietnamese. Journal of Social Psychology. 1971, 84, 183-190.

Mullins, C. J., Keith, J. B., and Riederich, L. D. Selection of foreign students for training in the United States Air Force. AFHRL-TR-68-111. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, November 1968.

Sinaiko, H. W. Verbal factors in human engineering: some cultural and psychological data. Ethnic Variables in Human Factors Engineering. (Ed) Alphonse Chapanis. Baltimore, MD: Johns Hopkins University Press, 1975.

TABLE 1

PERCENT DISTRIBUTION OF REFUGEES BY SEX (N=328)

	PERCENTAGE
MALE	82
FEMALE	18
TOTAL	100

TABLE 2

PERCENT DISTRIBUTION OF REFUGEES BY AGE (N=328)

	PERCENTAGE
18 to 22	36
23 to 27	36
28 to 32	16
33 to 35	12
TOTAL	100

TABLE 3

PERCENT DISTRIBUTION OF REFUGEES BY EDUCATION LEVEL (N=300)

	PERCENTAGE
NO EDUCATION	3
COMPLETED ELEMENTARY SCHOOL	12
COMPLETED SECONDARY SCHOOL	56
COMPLETED TRADE/TECH SCHOOL	6
COMPLETED UNIVERSITY	21
COMPLETED POST GRADUATE	2
TOTAL	100

TABLE 4

PERCENT DISTRIBUTION OF REFUGEES BY PRIOR MILITARY SERVICE (N=304)

	PERCENTAGE
PRIOR MILITARY SERVICE	71
NO PRIOR MILITARY SERVICE	29
TOTAL	100

TABLE 5

PERCENT DISTRIBUTION OF REFUGEES WITH PRIOR MILITARY SERVICE BY MILITARY OCCUPATIONAL TITLE (N=218)

	PERCENTAGE
ELECTRONICS	5
COMMUNICATIONS	6
MECHANICS	8
ADMINISTRATIVE	11
GENERAL TECHNICAL	8
GENERAL MILITARY DUTIES	62
TOTAL	100

NOTE: The military occupational titles referred to in this table were derived from a questionnaire item listing six occupational clusters commonly used to group military jobs. The examinees were asked to describe the military jobs they held while still on active duty in Southeast Asia. As indicated in the table, the vast majority of the refugees received only basic training while serving in the military. Factors contributing to their small amount of technical training are the relatively short length of service and enlisted status of the refugees.

TABLE 6

PERCENT DISTRIBUTION OF REFUGEES BY CIVILIAN OCCUPATION (N=143)

	PERCENTAGE
STUDENT	46
DRIVER	4
MEDICAL	2
COOK	4
TEACHER	8
MECHANICAL	6
SALES	9
CRAFTSMAN	19
MARITIME RELATED	2
TOTAL	100

NOTE: This distribution was obtained from a questionnaire item asking those refugees without prior military experience to write in the last civilian occupation in which they were employed. Categories were developed as the questionnaires were tallied. The large percentage declaring to be students reflects the young and single characteristics of the refugee population remaining in the camp.

TABLE 7

PERCENT DISTRIBUTION OF REFUGEES BY
ENGLISH LANGUAGE TRAINING (N=288)

SOURCE OF TRAINING	PERCENTAGE
HIGH SCHOOL	76
INDIANTOWN GAP MILITARY RESERVATION	14
MILITARY SERVICE	4
VIETNAMESE AMERICAN ASSOCIATION	5
UNIDENTIFIED	1
TOTAL	100

TABLE 8

PERCENT DISTRIBUTION OF REFUGEES BY FUTURE JOB PREFERENCE (N=265)

	PERCENTAGE
MILITARY SERVICE	4
CIVILIAN ELECTRONICS	22
CIVILIAN COMMUNICATIONS	6
CIVILIAN MOTOR MECHANICS	37
CIVILIAN ADMINISTRATIVE	15
CIVILIAN GENERAL TECHNICAL	16
TOTAL	100

NOTE: The job preference categories were derived from the six occupational clusters referred to in Table 5. The examinees were asked to select the one grouping in which they would prefer to be trained and employed. Briefly, Electronics included jobs as an electrician, computer technician or TV repairman. Communications included jobs as a radio or telegraph operator and radar technician. Motor Mechanics included a large category of jobs from truck driver to auto/aircraft repairman. The Clerical/Administrative grouping represented jobs as a secretary, typist, or bookkeeper. The General Technical cluster stood for jobs as a medical or dental technician, policeman or pharmacist. The two most popular categories related to jobs in the electronic and mechanical areas.

TABLE 9

DISTRIBUTION OF REFUGEES BY AGE AND EDUCATION LEVEL

SEX	EDUCATION LEVEL	18-22 YRS	23-27 YRS	28-32 YRS	33-35 YRS	TOTAL
MALE	NONE	1	3	1	1	6
FEMALE		1	0	1	2	4
MALE	ELEMENTARY	11	10	2	2	25
FEMALE		5	3	1	2	11
MALE	SECONDARY	50	60	22	14	146
FEMALE		8	4	5	3	20
MALE	TRADE/TECHNICAL	5	4	3	2	14
FEMALE		2	0	0	1	3
MALE	UNIVERSITY	13	19	10	5	47
FEMALE		8	7	2	0	17
MALE	POST-GRADUATE	0	1	4	1	6
FEMALE		0	0	0	1	1
MALE	TOTAL*	80	97	42	25	244
FEMALE		24	14	9	9	56

*Twenty-eight failed to indicate education level.

878

962

963

TABLE 10

PERCENT MALE REFUGEES ACHIEVING A MINIMUM ENGLISH LANGUAGE PROFICIENCY SCORE BY AGE AND EDUCATION LEVEL (N=244)

	AGE LEVEL				
EDUCATIONAL LEVEL	18 TO 22	23 TO 27	28 TO 32	33 TO 35	
NONE	0	0	0	0	
ELEMENTARY	0	1.03	0	0	
SECONDARY	2.50	5.15	2.38	8.0	
TRADE/TECH	0	0	2.38	4.0	
UNIVERSITY	2.50	3.09	9.52	12.0	
POST GRADUATE	0	1.03	2.38	0	
TOTAL	5	10.30	16.66	24.00	OVERALL 11.07

TABLE 11

PERCENT FEMALE REFUGEES ACHIEVING A MINIMUM ENGLISH LANGUAGE PROFICIENCY SCORE BY AGE AND EDUCATION LEVEL (N=56)

EDUCATIONAL LEVEL	AGE LEVEL				
	18 TO 22	23 TO 27	28 TO 32	33 TO 35	
NONE	0	0	0	0	
ELEMENTARY	0	0	0	0	
SECONDARY	0	0	11.11	0	
TRADE/TECH	0	0	0	0	
UNIVERSITY	8.33	14.29	0	0	
POST GRADUATE	0	0	0	11.11	
TOTAL	8.33	14.29	11.11	11.11	OVERALL 10.71

TABLE 12

CORRELATION COEFFICIENTS OF EACH DEMOGRAPHIC/BIOGRAPHIC QUESTIONNAIRE ITEM WITH ENGLISH COMPREHENSION LEVEL SCREENING TEST SCORES

QUESTIONNAIRE ITEM	COEFFICIENT OF CORRELATION
SEX	-.0080
AGE	.2138
COUNTRY OF BIRTH	-.0692
MARITAL STATUS	-.0749
EDUCATION LEVEL	.5111
PRIOR MILITARY SERVICE	.0406
BRANCH OF SERVICE	-.0429
LENGTH OF SERVICE	.0819
MILITARY OCCUPATION TITLE	-.0198
LENGTH OF MILITARY TRAINING	.0616
CIVILIAN OCCUPATION	-.0421
ENGLISH LANGUAGE TRAINING	.0709
LENGTH OF ENGLISH LANGUAGE TRAINING COURSE	.3252
FUTURE JOB PREFERENCE	.0042

TABLE 13

PERCENT REFUGEES ACHIEVING A MINIMUM ENGLISH

PROFICIENCY SCORE BY AGE AND PRIOR MILITARY SERVICE

PRIOR MILITARY SERVICE MALE (N=250)	AGE LEVEL 18 TO 22	23 TO 27	28 TO 32	33 TO 35	OVERALL PERCENTAGE
YES	0	10.00	17.50	23.08	10.38
NO	14.81	12.50	0	0	10.42

PRIOR MILITARY SERVICE FEMALE (N=54)	18 TO 22	23 TO 27	28 TO 32	33 TO 35	OVERALL PERCENTAGE
YES	0	0	0	0	0
NO	9.09	15.38	0	12.50	10.00

882

968

967

TABLE 14

RECOMMENDED DURATION OF ENGLISH LANGUAGE TRAINING IN WEEKS

IF ENGLISH COMPREHENSION LEVEL (ECL) IS:	TO GRADUATE WITH 60 ECL PLAN FOR:	65 ECL PLAN FOR:	70 ECL PLAN FOR:
24 OR LESS	23	26	30
25 TO 33	18	21	25
34 TO 42	14	17	21
43 TO 49	12	14	18
50 TO 54	10	12	15
55 TO 59	8	10	13
60 TO 64		8	11
65 TO 69			9

TABLE EXTRACTED FROM DLI PAMPHLET NUMBER 350.3

ECLST CUT-OFF SCORES FOR ENTRY OF FOREIGN MILITARY PERSONNEL INTO U.S. MILITARY TRAINING PROGRAMS.*

ARMY	60
NAVY	70
AIR FORCE	70
MARINE CORPS	70
COAST GUARD	70

*VERIFIED BY REPRESENTATIVES OF EACH SERVICE.

TABLE 15

PERCENT DISTRIBUTION OF MALE REFUGEES BY AGE SCORING 0-69
ON THE ENGLISH COMPREHENSION LEVEL SCREENING TEST (ECLST)

ECLST SCORES	AGE LEVEL				OVERALL PERCENTAGE
	18 TO 22	23 TO 27	28 TO 32	33 TO 35	
0 - 24	59	52	33	48	50%
25 - 33	23	21	14	12	19%
34 - 42	11	7	19	12	11%
43 - 49	3	6	14	0	6%
50 - 54	0	4	2	4	2%
55 - 59	0	0	2	4	.5%
60 - 64	0	0	0	8	.5%
65 - 69	0	4	2	0	2%
TOTAL	96	94	86	88	91%

884

TABLE 16

PERCENT DISTRIBUTION OF FEMALE REFUGEES BY AGE SCORING 0-69
ON THE ENGLISH COMPREHENSION LEVEL SCREENING TEST (ECLST)

ECLST SCORES	AGE LEVEL 18 TO 22	23 TO 27	28 TO 32	33 TO 35	OVERALL PERCENTAGE
0 - 24	46	43	56	56	48%
25-- 33	29	7	0	22	18%
34 - 42	8	21	11	11	13%
43 - 49	8	14	22	0	11%
50 - 54	0	0	0	0	0%
55 - 59	0	0	0	0	0%
60 - 64	0	0	0	11	2%
65 - 69	8	7	0	0	5%
TOTAL	99	92	89	100	97%

885

TABLE 17

MINIMUM MENTAL REQUIREMENTS FOR ENLISTMENT INTO THE ARMED FORCES, MALES AND FEMALES, 18 YEARS OF AGE, HIGH SCHOOL GRADUATES

	Minimum Qualifying Test Score Armed Forces Qualification Test/ Women's Selection Test		Service Unique Test Score
	AFQT	WST	ASVAB COMPOSITES
USA*			
MALE	16		90 on one of the following Army aptitude composites: Electronics (EL), General Maintenance (GM), Motor Mechanics (MM), Clerical (CL), General Technical (GT), Surveillance and Communications (SC).
FEMALE		59	N/A
USN			
MALE	21		N/A
FEMALE	21		N/A

TABLE 17 (Continued)

	AFQT	WST
USAF		
MALE	31	45 on General (G) of Air Force aptitude composites plus the sum of composites (Mechanical + Administrative + General + Electronics) equaling 170
FEMALE	31	"
USMC		
MALE	21	85 on General Technical
FEMALE	61	96 on General Technical
USCG		
MALE	25	N/A
FEMALE	25	N/A

*Applicants cannot be enlisted in Army solely on ASVAB-2. Once applicant attains minimum AFQT/WST score and minimum subtest score (if applicable), then males must take the Army Classification Battery (ACB) Book 2 and females must take "Classification and Inventory" of ACB Book 2.

887

TABLE 18

ASVAB TESTED REFUGEES QUALIFYING FOR THE ARMED FORCES BY BRANCH OF SERVICE AND SEX

	BRANCH OF SERVICE					
SEX	USA	USN	USAF	USMC	USCG	ALL BRANCHES
MALES	24	24	12	17	24	11
% FROM ORIGINAL SAMPLE (N=269)	9%	9%	4%	6%	9%	4%
FEMALES	1	4	1	1	4	1
% FROM ORIGINAL SAMPLE (N=59)	2%	7%	2%	2%	7%	2%
TOTAL	24	27	12	17	27	12
%ORIGINAL SAMPLE ENLISTMENT QUALIFIED (N=328)	7%	8%	4%	5%	8%	4%

888

TABLE 19

ASVAB TESTED REFUGEES BY
DERIVED AFQT MENTAL CATEGORY AND SEX

SEX	I 93%-ABOVE	II 65%-92%	III 31%-64%	IV 10%-30%	V 0%-9%	TOTAL
MALES	0	6	15	4	1	26
%	0.0%	23%	58%	15%	4%	100%
FEMALES	0	0	4	2	1	7
%	0.0%	0.0%	57%	29%	14%	100%
TOTAL	0	6	19	6	2	33
% OF TOTAL TESTED	0.0%	18%	58%	18%	6%	100%

CHARACTERISTICS OF SELF SELECTED JOB CONTENT[1]

Barry P. McFarland
Aeronautical Systems Division
James B. Carpenter
St Mary's University

ABSTRACT

As a part of a comprehensive job analysis performed on the Air Force's Medical Service Career Field a special job inventory was developed to determine if medical service corpsman could design a job consisting of sets of self selected tasks that were realistic in representing an Air Force need. These "ideal" jobs were to be related to the existing job types in the medical service career field.

An experimental job inventory consisting of 600 individual tasks was completed by 1207 Air Force medical service personnel who had previously participated in a job analysis survey. The experimental inventory requested the subject complete the inventory by identifying all tasks which would comprise their "ideal" job.

Comparisons were made with existing job types. The jobs types defined by the "ideal" job inventory has implications for job reengineering, job satisfaction, and job expectancy as it may impact recruiting, training, and retention of personnel.

1 This research was conducted while both authors were assigned to the Occupational and Manpower Research Division, Air Force Human Resources Laboratory, Lackland AFB TX.

I. INTRODUCTION

This study was part of a comprehensive experimental program developed in cooperation with the Nursing Resources Study Group appointed by the Air Force Surgeon General. The Nursing Resources Study Group was interested in Air Force-wide information on Air Force nurse and medical service specialities for use in determining current and future Air Force nurse requirements. This report describes the use of an "Ideal Job" scale that was developed and used to assist the study group in determining task trade-offs that could be made allowing for more effective use of both nurse and medical service personnel. Previous reports in this series include Archer (1973), McFarland (1973), McFarland (1974), McFarland (1976). They describe the job analysis performed on nurse and medical service personnel, the job inventories used, and the difficulty of jobs performed.

Job Analysis Methodology

Over the past 10-15 years the Occupational and Manpower Research Division of the Air Force Human Resources Laboratory has conducted extensive research in the areas of job analysis and job inventory methodology. The job analysis procedure they developed has proven to be a vital tool for the effective operation of the Air Force's personnel management system. The job analysis system has served as the basis for modification of the personnel classification structure, personnel selection techniques, evaluation of job incumbents performance, and for determination of appropriate job reengineering actions. Additionally it has been used extensively to determine training requirements as an integral part of the Air Force's Instructional Systems Development program (see Christal 1974 for review). The basic approach and rationale for the Air Force's approach to job analysis has been comprehensively reported in a series of reports dating from the early sixties (Morsh, Madden and Christal 1961; Morsh 1962; Archer and Fruchter 1963; Morsh and Archer 1967; Christal 1974).

As suggested above the job analysis methodology is a dynamic process intended to improve and meet new personnel management requirements in a timely manner. An integral part of this dynamic process has been the development of several different scales that are amenable to the job inventory format. These scales include criticality scales time spent scales, and training need scales. Each of these were developed as a result of specific personnel management requirements.

This paper describes the use of a new scale developed specifically to look at job redesign for the medical service career field. The scale is called the "Ideal Job Scale."

II. PROCEDURE

The Ideal Job Scale

All persons in the sample were requested to describe their "Ideal Job" by selecting a set of individual tasks. The scale used was a seven point scale which required the job incumbents to rate the relative amount of time they would spend on each task as part of their ideal job. The following restrictions were placed on this ideal job:

1. The job had to serve some useful purpose within a Air Force medical facility.

2. The job had to be performed with only minimal additional training. Aside from these two restrictions, the job incumbents could specify any number of tasks and distribute their relative time to these tasks in any way to define their ideal job.

The Job Inventory

The tasks in the job inventory were identical to those previously reported by McFarland (1973). The inventory consisted of 600 tasks group under 11 different duty headings. The job inventory was mailed to approximately 1800 medical service personnel who had completed a previous medical service job inventory and were still in the Air Force. Of the 1800 only those who had not changed jobs or locations were included in the sample for analysis.

III. DATA ANALYSIS

The first step in the analysis,was to convert each individual's relative time-spent responses (1-7 scale) to percent time ratings. To obtain the percent time ratings, all of an incumbent's time-spent ratings are summed and the total is assumed to represent 100 percent of his time spent on the job. Each rating is divided by the total and the quotient multiplied by 100 to give a percent time spent estimate on each task. For the purpose of organizing jobs into similar units of work, an automated job-clustering computer program was used. This hierarchical-grouping program is a basic part of the job analysis system. The program compares each individual with every other individual in the sample in terms of percent time spent on each and every task in the inventory. The two persons with the most similar jobs are located and combined to form a group with a composite job description. In successive stages, the program adds other members to this group or forms new groups based on similarity in the percent of time spent on tasks description. This procedure is continued until all individuals and groups are combined to form a single group. At each stage of the grouping process

an index of homogeneity is calculated. This index, percentage of work overlap of group members. The index is an estimate of the overlap of work that would be expected if a member of the group was randomly reassigned to a job in that same group.

A diagram of the results of the hierarchical grouping procedures are included in Figure 1. Each group is identified by a group number indicating the stage in the program during which the group was formed, and number of subjects in the group. For example, GRP 001 was formed at group stage 1, the last stage of the grouping program. Since it is the last stage, this is equivalent to the total sample.

For those designated groups,the raw data from both the background information and task listing were converted into a form more readily interpretable for identifying job types. Reports obtained included background variables with descriptive statistics for quantifiable data and lists of tasks with percent members performing, and average percent time values for each task. Additional reports, identifying the differences in those tasks performed by the major job types and the identification of the prime tasks performed by each job type, are available upon request.

IV. RESULTS

A description of the 25 tasks that comprized the largest amount of time spent by medical service corpsmen in their ideal job is included in table 1. As can readily be seen from inspection of these tasks they include mostly tasks associated with either emergency room functions or triage functions. Table 2 shows the comparision between tasks performed in the ideal job versus those tasks currently performed. Note again the emphasis on the emergency room and triage functions versus the administrative and house keeping task functions. What is particularly interesting however, is the fact that even in their ideal job a number of medical service corpsmen included some of the administrative task, albeit not to the same extent as they presently perform these tasks.

To test the hypothesis that jobs identified by the ideal job scale were in fact realistic jobs, a hierarchial cluster analysis was used to compare jobs presently performed with the ideal jobs. These results are shown in figure 1. As can be seen there are a number of clusters in which the ideal job did cluster with jobs currently being performed. This suggests that for many, the ideal job is currently available within the Air Force. Table 3 shows the number of tasks representing the cumulative percent time spent by each group. This serves as a rough estimate of homogeneity. A brief description of each of these clusters follows:

Group 456

As shown in figure 1 this group consists of 258 current job descriptions and 83 ideal jobs. The group is representative of emergency room or

TABLE 1. PRIMARY TASKS IDENTIFIED BY IDEAL JOB SCALE

D-TSK	DUTY/TASK TITLE	CUMULATIVE SUM OF AVERAGE PERCENT TIME SPENT BY ALL MEMBERS	AVERAGE PERCENT TIME SPENT BY ALL MEMBERS	AVERAGE PERCENT TIME SPENT BY MEMBERS PERFORMING	PERCENT OF MEMBERS PERFORMING
G 27	SUTURE LACERATIONS (502)	57.58	1.35	.78	.78
B 16	ADMINISTER INTRAMUSCULAR MEDICATIONS (165)	65.62	1.02	.67	1.45
B 12	ADMINISTER FIRST AID (161)	61.56	1.07	.66	2.11
E 63	TAKE ELECTROCARDIOGRAPH (EKG) TRACINGS (460)	54.68	1.17	.64	2.75
G 22	REMOVE SUTURES (497)	58.49	1.09	.64	3.38
B 181	TAKE AND RECORD BLOOD PRESSURES (330)	54.02	1.14	.62	4.00
A 18	COORDINATE WITH PHYSICIAN REGARDING PATIENT CARE (18)	54.02	1.11	.60	4.60
A 30	DEVELOP OR IMPROVE WORK METHODS OR PROCEDURES (30)	50.21	1.17	.59	5.19
B 35	APPLY BANDAGES (184)	59.40	.98	.58	5.77
B 182	TAKE AND RECORD PULSES, TEMPERATURES AND RESPIRATIONS (331)	52.44	1.10	.57	6.35
A 2	ANSWER TELEPHONE CALLS FOR OR FROM PATIENTS OR HOSPITAL STAFF MEMBERS (2)	35.87	1.53	.55	6.90
B 15	ADMINISTER INTRADERMAL MEDICATIONS (164)	59.15	.92	.54	7.44
B 70	CHANGE DRESSINGS (219)	56.59	.96	.54	7.99
G 18	PERFORM MINOR SURGERY (493)	41.84	1.26	.53	8.51
G 2	ADMINISTER IMMUNIZATIONS (477)	42.75	1.22	.52	9.04
A 66	LABEL SPECIMENS AND FILL OUT LABORATORY SPECIMEN REQUEST FORMS (66)	46.81	1.08	.51	9.54
A 145	SUPERVISE 902X0 OR 902X2 PERSONNEL (145)	38.44	1.30	.50	10.04
B 27	ADMINISTER SUBCUTANEOUS MEDICATIONS (176)	57.33	.87	.50	10.54
B 19	ADMINISTER IPPB (INTERMITTANT POSITIVE PRESSURE BREATHING) THERAPY (168)	55.18	.90	.50	11.04
B 62	ASSIST WITH SPINAL PUNCTURE (211)	57.66	.85	.49	11.53
A 29	DETERMINE WORK PRIORITIES (29)	42.34	1.13	.48	12.00
B 176	SET UP OR USE OXYGEN EQUIPMENT (325)	52.94	.86	.46	12.46
B 25	ADMINISTER ORAL MEDICATIONS (174)	55.18	.82	.46	12.92
B 32	ADMINISTER WOUND IRRIGATIONS (181)	52.86	.85	.45	13.37
A 20	COORDINATE WITH WARDMASTER (NCOIC) REGARDING PATIENT CARE (20)	41.43	1.08	.45	13.82
B 61	ASSIST WITH CUTDOWNS (210)	51.12	.85	.44	14.25
B 155	PREPARE PATIENTS FOR PHYSICIAN EXAMINATION OR TREATMENT (304)	45.32	.95	.43	14.69
B 17	ADMINISTER INTRAVENOUS INFUSIONS (166)	50.12	.82	.41	15.10
B 111	MAKE WARD ROUNDS WITH PHYSICIANS (260)	44.08	.93	.41	15.51
G 28	SUTURE SURGICAL INCISIONS (503)	34.71	1.18	.41	15.92

TABLE 2. DIFFERENCE BETWEEN IDEAL JOB AND PRESENT JOB DESCRIPTION

PERCENT MEMBERS PERFORMING---DIFFERENCE, SPC001 MINUS SPC002
Ideal Job PERCENT MEMBERS PERFORMING
Present Job PERCENT MEMBERS PERFORMING

D-TSK		TASK TITLE			
A	2	ANSWER TELEPHONE CALLS FOR OR FROM PATIENTS OR HOSPITAL STAFF MEMBERS (2)	80.93	35.87	45.05
B	133	PERFORM GENERAL HOUSEKEEPING DUTIES (282)	65.34	25.27	40.07
B	181	TAKE AND RECORD BLOOD PRESSURES (330)	83.50	54.02	29.48
B	182	TAKE AND RECORD PULSES, TEMPERATURES AND RESPIRATIONS (331)	81.09	52.44	28.65
B	71	CLEAN AND MAINTAIN EQUIPMENT (220)	73.38	41.84	31.54
A	101	PREPARE CONSULTATION REQUESTS OR LAB SLIPS (101)	63.35	33.72	29.63
B	88	FOLD OR COUNT LINEN (237)	61.11	26.35	34.76
B	77	COLLECT AND LABEL SPECIMENS SUCH AS URINE, FECES OR SPUTUM FROM PATIENTS (226)	64.51	37.61	26.90
B	104	MAKE BEDS OTHER THAN POSTOPERATIVE OR RECOVERY (253)	52.82	25.77	27.05
B	74	CLEAN PATIENT CARE UNIT (223)	50.66	26.26	24.40
B	76	CLEAN WARD UTILITY ROOM (225)	49.00	23.45	25.56
A	66	LABEL SPECIMENS AND FILL OUT LABORATORY SPECIMEN REQUEST FORMS (66)	70.07	46.81	23.26
B	73	CLEAN MINOR SURGERY OR EXAMINATION ROOMS (222)	51.58	30.32	21.25
A	6	ASSIST PATIENTS IN PLACING TELEPHONE CALLS (6)	61.77	27.26	34.52
B	78	COLLECT FOOD TRAYS OR SERVING UNITS (227)	47.10	24.36	22.74
B	81	DISPOSE OF CONTAMINATED MATERIALS (230)	58.71	27.01	31.70
B	83	DRAW BLOOD FOR LABORATORY TESTS (232)	27.45	46.15	-18.70
B	63	ASSIST WITH STERNAL PUNCTURE (212)	13.60	36.70	-23.10
B	64	ASSIST WITH THORACENTESIS OR PARACENTESIS (213)	23.47	43.33	-19.86
B	57	ASSIST PHYSICIAN IN MAKING DIFFERENTIAL DIAGNOSIS (206)	16.75	35.21	-18.46
C	2	ASSIST SURGEON DURING OPERATION BY RETRACTING OR SPONGING (343)	13.18	32.73	-19.54
B	17	ADMINISTER INTRAVENOUS INFUSIONS (166)	29.44	50.12	-20.69
B	80	DELIVER BABY IN ABSENCE OF PHYSICIAN (229)	13.93	36.29	-22.36
B	59	ASSIST WITH BIOPSIES (208)	21.72	40.18	-18.46
B	18	ADMINISTER INTRAVENOUS MEDICATIONS (167)	26.53	48.14	-21.60
E	25	INTERPRET ELECTROCARDIOGRAPHIC TRACINGS (422)	10.45	29.08	-18.63
E	11	CONDUCT SICK CALL IN ABSENCE OF PHYSICIAN (408)	10.61	31.57	-20.95
B	111	MAKE WARD ROUNDS WITH PHYSICIANS (260)	23.71	44.08	-20.36
B	61	ASSIST WITH CUTDOWNS (210)	28.61	51.12	-22.51
	28	SUTURE SURGICAL INCISIONS (503)	14.68	34.71	-20.04
	18	PERFORM MINOR SURGERY (493)	20.23	41.84	-21.61
	27	SUTURE LACERATIONS (502)	33.75	57.58	-23.83

dispensory type job descriptions. The following 9 tasks represent the tasks performed by the largest percent of job incumbents and account for the greatest relative percent of time spent.

1. Apply bandages.
2. Change dressings.
3. Administer first aid.
4. Take and record blood pressures.
5. Clean and maintain equipment.
6. Take and record pulses, temperatures and respirations.
7. Suture lacerations.
8. Remove sutures.
9. Administer intramuscular medications.

Group 566

This group consisted almost entirely of ideal job descriptions. As shown in table 3 this group was significantly more heterogeneous then group 456 in that substantially more tasks account for the same percent of time. Like group 456 the tasks were representative of emergency room job types as shown by the following representative tasks from this group:

1. Suture lacerations.
2. Administer first aid.
3. Administer interamuscular medications.
4. Assist in spinal puncture.
5. Assist with cut downs.
6. Apply bandages.
7. Administer wound irrigations.
8. Administer intradermal medications.
9. Remove sutures.

TABLE 3. DESCRIPTIVE TASK DATA FOR MAJOR JOB CLUSTERS

Group Number	Number of Members	Number of Tasks Performed as a Percent of Time	
		25%	95%
456	342	29	331
566	128	47	408
561	434	58	468
399	423	25	309
185	133	27	370
182	144	18	286
117	80	14	256
47	263	25	341
34	123	12	264
Ideal Job	1207	54	286
Present Job	1207	38	409

Group 561

This group consists of an almost equal number of present and ideal job descriptions. The tasks were representative of jobs found in a medical/surgery ward. Representative tasks of this job type include:

1. Take and record pulses, temperatures and respirations.
2. Take and record blood pressures.
3. Apply bandages.
4. Collect and label specimens.
5. Change dressings.
6. Measure and record intake and output.
7. Ambulate patients.
8. Administer IPPB.
9. Administer oral medications.

Group 399

This group is representative of the typical ward corpsmen job descriptions. The relative level of difficulty for this job cluster is substantially less then any described thus far, as evidenced by the following representative lists of tasks:

1. Take and record pulses, temperatures and respirations.
2. Take and record blood pressures.
3. Collect food trays or serving carts.
4. Make beds.
5. Clean patient unit.
6. Clean ward utility room.
7. Administer bed pans and urinals.
8. Clean and maintain equipment.
9. Collect and label specimens.

Group 185

This group like group 566 consists almost entirely of ideal jobs. The tasks representing this job are very difficult and the job type could best be described as a "junior physicans" job type. Representative tasks from this job include:

1. Suture lacerations.
2. Administer first aid.
3. Administer intramuscular medication.
4. Assist with cut downs.
5. Administer intravenous infusions.
6. Assist with spinal puncture.
7. Perform minor surgery.
8. Apply bandages.
9. Administer intradermal medications.

Group 182

This job type containing both ideal job and current job descriptions is typical of the outpatient clinic job types. Representative tasks in this job type are:

1. Answer telephone.
2. Label specimens.
3. Take and record blood pressures.
4. Prepare consultation requests or lab slips.
7. Clean and maintain equipment.
8. Prepare infant ID bands or foot prints.
9. Collect food trays.

Group 47

This group containing both ideal job and current job descriptions is an administrative and supervisory job type as shown by the following set of representative tasks:

1. Supervise 902 x 0 or 902 x 2 personnel.
2. Counsel personnel on performance.
3. Counsel personnel on personal problems.
4. Evaluate duty performance of personnel.
5. Coordinate work activities with other sections.
6. Develop or improve work methods or procedures.
7. Determine work priorities.
8. Administer performance examinations.
9. Prepare or endorse airmen performance reports.

Group 34

The last job typing to be described also consists of a mixture of ideal job descriptions and current job descriptions. The tasks describe a job in a physical exam clinic. Representative tasks include:

1. Prepare consultation or lab slips.
2. Take and record blood pressure.
3. Answer telephone.
4. Take and record pulses, temperatures and respirations.
5. Label specimens.
6. Administer hearing tests.
7. Administer vision tests.
8. Order x-rays for patients.
9. Collect and maintain health statistics.

IV. DISCUSSION AND CONCLUSIONS

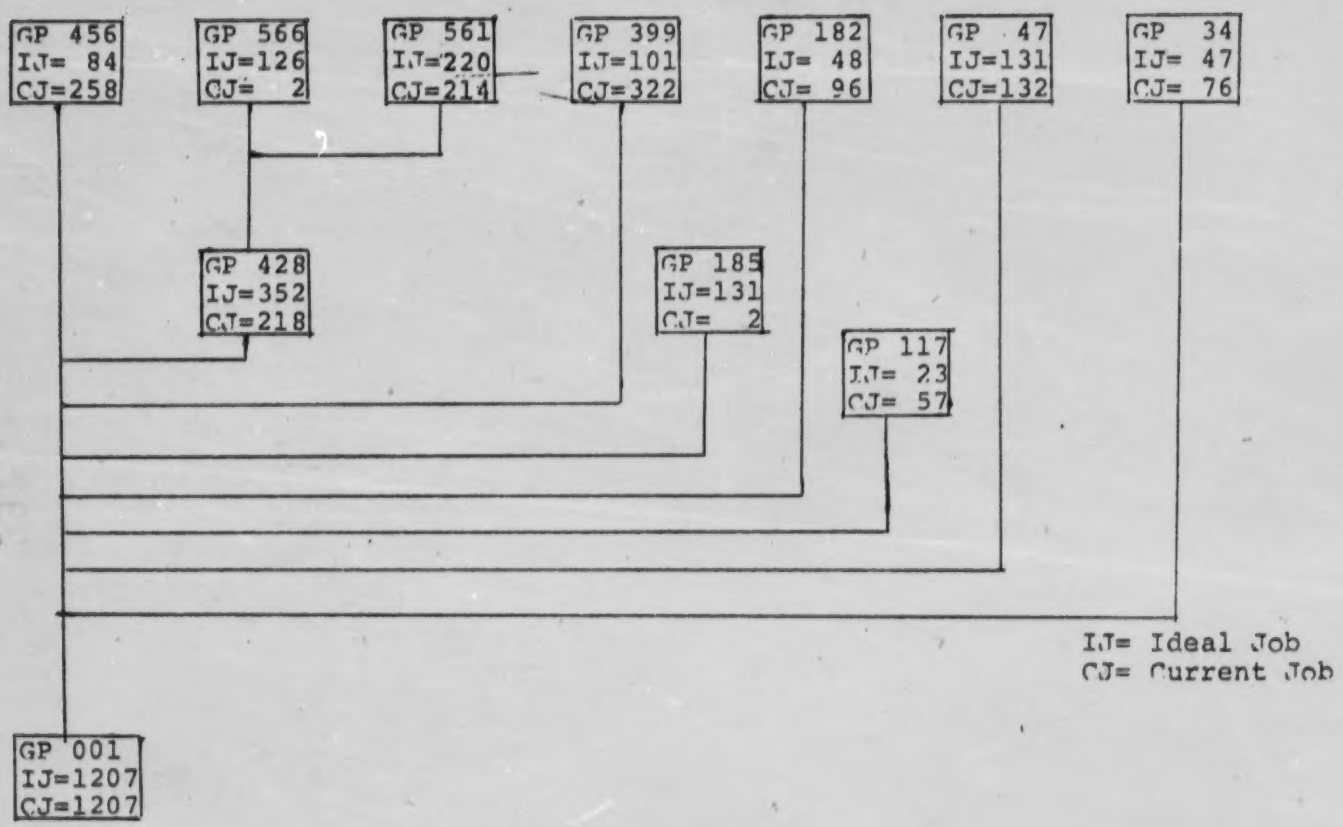
As shown in the task descriptions and the cluster analysis the medical service corpsmen did consistently design realistic job description using the ideal job scale. Of greater interest is the fact that where the job descriptions were different from the existing job types, the jobs were consistently more difficult than those jobs being currently performed. It should also be noted that the ideal job descriptions consistently included tasks that involved more direct contact with the patients and substantially less administrative work. Both of these findings are significant for job redesign in this career field. It appears, first that the medical service corpsman is seeking greater responsibility in terms of a more difficult job with more patient contact and secondly, they have been trained and are capable of performing these tasks as evidenced by the fact that these tasks are currently being performed in the career field.

All results obtained to date for the job performance of the medical service career field strongly indicate that this is a highly motivated and extremely competent group of personnel. It is felt that as the nursing corps is given more responsibility in such programs as the pediatric nurse practitioner, nurse mid-wife and triage nurse programs, that the medical service corpsman can and is willing to accept the additional responsibilities that may be left by the nurse.

The experience with the ideal job scale has suggested a number of other possible uses of the scale in the area of job analysis. First it would appear that the ideal job could be useful in identifying tasks that are potential job satisfiers. This might permit investigators to redesign jobs in which the job content is causing job dissatisfaction. In this study, however there were no differences found between job satisfaction in the current job when compared to the expressed satisfaction for the ideal job. This finding is not surprising since the medical service personnel express a high level of satisfaction with their present jobs.

A second area in which this ideal job scale may lend itself for investigation is in the area of job expectancy, with potential impacts on both training and recruiting. Since all personnel were required to design their ideal job within the constraints of minimal additional training, some insight might be gained by looking at the jobs they defined as a description of the tasks for which they were trained and subsequently the tasks they expected to perform upon completion of training. Again this is not much of a problem in the medical service career area, but in other career fields in which job content is not consistent with training the ideal job scale might provide some critical insight into the problem.

Figure 1. Cluster Analysis Results



REFERENCES

Air Force Manual 35-2. Occupational analysis precedures for conduction occupational surveys and Air Force specialty evaluations. Washington, DC; Department of the Air Force, 9 July 1968.

Archer, W. B. Job analysis of the nursing career field. Unpublished report 1973.

Archer W. B. & Fruchter, D. A. The construction review and administration of Air Force job inventories. PRL-TDR-63-21 Lackland AFB TX; Personnel Research Laboratory Aerospace Medical Division, August 1963.

Christal, R. E. The United States Air Force occupational research project. AFHRL-TR-73-75 Lackland AFB TX; Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

McFarland, B. P. Job analysis of the medical service career field. AFHRL-TR-73-36. Lackland AFB TX; Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

McFarland, B. P. A comparison of task difficulty ratings made by nurses and medical service corpsman. AFHRL-TR-74-7. Lackland AFB TX; Occupational Research Division, Air Force Human Resources Laboratory, March 1974.

McFarland, B. P. A comparative analysis of nurse and medical service personnel, AFHRL TR-76-xx, Lackland AFB TX; Occupational and Manpower Research Division, Air Force Human Resources Laboratory (in press).

Morsh, J. E. Research findings obtained with a task inventory method of job analysis, American Psychologist, 1962, 17, 395.

Morsh, J. E. & Archer, W. B. Procedural guide for conduction occupational surveys in the United States Air Force. PRL-TR-67-11. Lackland AFB TX; Personnel Research Laboratory, Aerospace Medical Division, September 1967.

Morsh, J. E., Madden, J. M. & Christal, R. E. Job analysis in the United States Air Force. WADD-TR-61-113. Lackland AFB TX; Personnel Research Laboratory, Wright Air Development Division, February 1961.

EVALUATION OF G-SEAT CUEING IN LARGE BODIED AIRCRAFT

John C. Griffin
Barry P. McFarland
Oak H. DeBerg
Aeronautical Systems Division
Wright-Patterson AFB, Ohio 45433

ABSTRACT

The g-seat, a haptic cueing device for flight simulators, has been proposed as a supplemental cueing device for large cabin aircraft, or even as a replacement for the platform motion systems now in use on most modern aircraft simulators. Devices of this type have been in use for several years, but evaluated primarily in simulators of fighter type aircraft, whose flight regime contain high "G" loads. To evaluate the cueing effectiveness of a g-seat in large bodied aircraft simulators, a joint Air Force/NASA study was conducted.

The purpose of the study was to determine the quality of the transfer of training for four different cueing configurations. These were formed by using the four combinations of platform motion and G-seat motion. The testing was conducted in NASA's Flight Simulator for Advanced Aircraft, and the subjects were Air Force pilots.

In this paper, an initial comparison of the four systems is made. Limitations in existing G-seats are discussed with respect to the tasks analyzed, and recommendations for future research are made. The effects of platform motion cueing in these situations are also discussed.

I. INTRODUCTION

In recent years the advancements in aircraft simulator technology have been enormous. New devices are continually being placed on the market with the aim of improving the verasimility of the machine. A problem confronting those who must purchase these simulators is the determination of which of these new devices is useful and necessary and should be incorporated in the design. It is not enough for the device to make the simulator seem more like the aircraft. Cost constraints require that it aid in the function of the machine, typically the training of the student. Because of the speed with which the simulator field has grown, there is often a very limited amount of behavioral data available on new devices prior to the production decision. It has become necessary to evaluate the affects of some of these new devices to determine their real contributions to training programs.

Attempting to isolate the affects of one device from all others is an extremely difficult problem. Quite often the affects of one type of cue will be greatly effected by the presence or absence of another. In many cases cue interdependency leads to a synergestic effect, and the evaluation of the one cue in isolation is not only difficult but can produce misleading results. Also, the type of task being performed is a major factor to be considered in evaluating the worth of a cueing device. Obviously different tasks require different information be given to the pilot and while a certain cue may be of no use in one situation, it may be of great value in another. These problems make subsystem selection very argumental.

Motion cueing, in particular, has been the subject of a great deal of controversy. The question of what type of motion cueing is necessary to sufficiently train a pilot has become a highly sensitive and emotional issue. Authorities in the field are devided on the questions regarding the value of a platform motion system, the necessity of haptic cues, and the degree of fidelity needed in a visual system to eliminate a need for motion. Dispite all the controversy, very little factual data has been collected on the actual effects motion cueing has on the learning processes of pilots.

There is little argument that the cost of operating large bodied aircraft is high, and a great deal of money can be saved in the utilization of simulators for both training and flight proficiency. It is therefore desirable to obtain some information on the effects of motion systems on training in these large bodied aircraft simulators. To this end, a study sponsored jointly by the United States Air Force and the National Aeronautics and Space Administration (NASA) was conducted to evaluate the cueing effectiveness of a device known as the g-seat for certain tasks in a simulation of the KC-135 aircraft.

The g-seat is designed to provide stimuli for sustained acceleration cues. Through the use of seat movements, stimuli recognized by elements of the haptic sensory system are provided to the pilot which coincide with the simulator's condition. Devices of this type have been in use for several years. They have been used as supplemental cueing devices, and in some cases, as replacements for platform motion systems. This use, however, has been confined largely to fighter type aircraft, whose flight regimes contain high "G" loads.

The purpose of this study was to analize the effects of incorporating a g-seat in a simulator for large bodied aircraft, where the "G" loads seldom vary by more than ± 1. Because of simulator requirements, it has become a necessity to determine if the g-seat aids in training, if it adds to or detracts from the cues of the standard platform motion systems, and if it could be used as a replacement for such systems. The interaction of this device with the platform motion system necessitated that the affects of motion has on the transfer of training also be considered.

II. PROCEDURE

The transfer of training study consisted of three phases; familiarization, training and criterion. Phase 1, the familiarization, briefed the subjects on their duties, and allowed them to become familiar with the full capabilities of the simulator and g-seat being used for the study. In this phase each subject was permitted to fly the simulator for approximately 15 minutes to get the feel of the machine and learn its idiosyncrasies.

Phase 2 was training. In this phase the subjects were divided into four groups of equal size, and each group was given a different training system. These systems were used to teach each group of subjects an identical series of flight manuvers, or tasks. The four training systems were:

1. No motion, no g-seat.
2. No platform motion, g-seat.
3. Restricted platform motion, no g-seat.
4. Restricted platform motion, g-seat.

Each subject was trained on his assigned system until he became proficient in the given tasks.

The subjects then transitioned to phase 3 of the study. In this phase all of the subjects flew the tasks utilizing the full capabilities of the simulator. An assessment of the training effectiveness of each system was then obtained by comparing the scores of the four groups during this phase. Each task was flown twice by each subject, once with the g-seat on and once with it off. The order of presentation was varied from subject to subject, but an equal number had the g-seat first and last.

Apparatus

The simulator used for this study was the Flight Simulator for Advanced Aircraft (FSAA), an engineering research simulator located at the NASA Ames Research Center. This system was selected because of its outstanding ability to generate high fidelity motion cues for yaw, roll, and lateral movement in the simulation of the KC-135 aircraft. Because of the physical capabilities of this system (see table 1), the simulator is sufficient to be regarded as comparable to the aircraft for the tasks analyzed.

In order for the training to be similar to that possible in current USAF simulators, the platform motion system of the FSAA was restricted, during the training phase, to represent that of a typical "6 post" system. This was done through software modifications, using the concepts developed at the NASA Langley Research Center (ref 1), and the NASA Ames Research Center (ref 2).

The g-seat employed in this study was developed by the Singer Simulation Products Division for the USAF Advanced Simulator for Undergraduate Pilot Training (ASUPT). It consisted of 16 air cells, forming the seat cushion, 9 back air cells, and 3 thigh panel air cells for each thigh. The seat also was equipped with a pneumatically actuated lap belt. The surface positions of these cells and the belt varied independently under computer control in order to provide heptic system motion cues. (Ref 3,4,5.)

The visual presentation was generated by a Redifon model board visual system. This is a 6-degree of freedom system in which a color television camera is mounted on a gantry and moves relative to the fixed-model board. The scene is presented to the pilot, through collemated optics, on a cathode-ray tube situated in the forward cockpit window. The viewing angles at the pilot's eye are 36 degrees vertically and 48 degrees horizontally. Audio cues were also provided for the pilot in the form of engine noise, which varied with the engine performance.

The drive logic for the FSAA was provided by a Sigma 8 digital computer. This computer also was used for the coordination of the g-seat drive with the platform motion system, and for the collection of data.

TABLE 1

FSAA SPECIFICATIONS

Parameter	Dimensions	Acceleration
Roll	± 45 degrees	4 rad/sec^2
Yaw	± 30 degrees	2 rad/sec^2
Pitch	± 22 degrees	2 rad/sec^2
Lateral	± 50 ft	15 ft/sec^2
Vertical	± 5 ft	15 ft/sec^2
Longitudinal	± 4 ft	10 ft/sec^2

Tasks

Three types of tasks were performed by each of the subject pilots during the study. The first of these was a takeoff and straight-out departure under no wind conditions. At some point between lift-off and 1000 ft one of the outboard engines was failed. The pilots' job was to recover the aircraft and continue the departure as quickly and accurately as possible. The pilots were briefed before hand that an engine failure would occur on most trials, but the altitude and engine number were not revealed. This task was selected because of the large yaw cue provided in a sudden manner. The sudden, violent accelerations due to the engine loss made this a potentially relevant task for requiring motion cues.

The second task the pilots were trained to perform was very basic. It consisted of two 180 degree turns. The first was climbing from 500 ft at 1000 ft per minute, and the second descending from 1200 ft at the same rate. The directions of the turns was varied from trial to trial. This task was performed under "IFR" conditions. The visual scene was a gray field, simular to flying in clouds. The pilots were instructed to keep the vertical velocity, airspeed, and bank angle as constant as possible. Because of its simplicity and the pilots familiarization with it, it was assumed this task would not require many motion cues in order to be learned or performed. The elimination of the visual cues further reduced the input given to the pilots.

Task 3 was a landing. The pilot was instructed to intercept the glide slope at 1000 ft and using raw data fly as close to glide slope and localizer as possible until touchdown. The difficulty of this task was magnified by turbulent wind conditions, and the interjection of a wind shear at some point along the path. The wind shear varied from trial to trial in altitude, direction and magnitude. Again, the pilots were aware a wind shear might occur, but the details were not revealed. During this task it was assumed motion might play a very important role in achieving the best performance. The rarity of such winds also allowed for the possibility of more training to occur in this task than in task 2.

The same sequence of tasks and trials was given to each pilot, regardless of which training system he used. However, the pilots may have finished their trials at different points in the sequence, depending on the speed with which they became proficient.

Subject

The test subjects for this study were 48 US Air Force KC-135 aircraft commanders. Each of the four training systems was composed of 12 of these pilots. The pilots were further grouped as either experienced or inexperienced, with an equal number in each system. The average number of hours the experienced pilots had as KC-135 commanders was 1440, and the inexperienced averaged just 280 hours as commander.

III. RESULTS

The results presented in this paper are preliminary and should be treated as such. To avoid possible misinterpretation of the results we will present scatter plots of some of the raw data and make comments as to the trends we see emerging from it. The data presented is merely a representative sample of a much larger group of variables.

Task 1

Figures 1, 2, and 3 show the scatter plots of variables SUMI, H+V, and T_1 versus training system group for task 1, the outboard engine failure on takeoff. These variables are defined as follows:

$$\text{SUMI} = \int_{t_o}^{t_e} |r_B| \, dt + \int_{t_o}^{t_e} |p_B| \, dt + \int_{t_o}^{t_e} |q_B| \, dt$$

$$\text{H+V} = \int_{t_o}^{te} \left| \frac{\dot{h} - \dot{h}_{nom}}{\dot{h}_{nom}} \right| dt \Big/ (t_e - t_o) + \int_{t_o}^{t_e} \left| \frac{V_{rw} - V_{nom}}{V_{nom}} \right| dt \Big/ (t_e - t_o)$$

$$T_1 = t_{\partial_r} - t_o$$

where t_o = start time of manuver

t_e = end time of manuver

r_B = roll rate

p_B = pitch rate

q_B = yaw rate

$\dot{h}$ = altitude change (ft/sec)

t_{∂_r} = time of control input after engine failure

V_{rw} = relative wind (ft/sec)

Figure 1 indicates all groups kept the simulator under control, and except for one subject in group 2 who crashed, no one entered a "dutch roll." Groups 1 and 3, however, had much less variability. In contrast, figures 2 and 3 show that for the control of the optimal rate of climb and reaction time to the engine failure, groups 1 and 3 performed better than groups 2 and 4. It should be reiterated that both groups 2 and 4 were trained with a g-seat cueing device, and 1 and 3 were not.

Task 2

Figure 4 graphs the amount of power the pilot input in the controls. Power was defined as:

$$\text{Total Power} = \int_{t_o}^{t_e} \partial_w^2 \, dt \Big/ (t_e - t_o) + \int_{t_o}^{t_e} \partial_c^2 \, dt \Big/ (t_e - t_o) + \int_{t_o}^{t_e} \partial_p^2 \, dt \Big/ (t_e - t_o)$$

where ∂_w = number of wheel control inputs

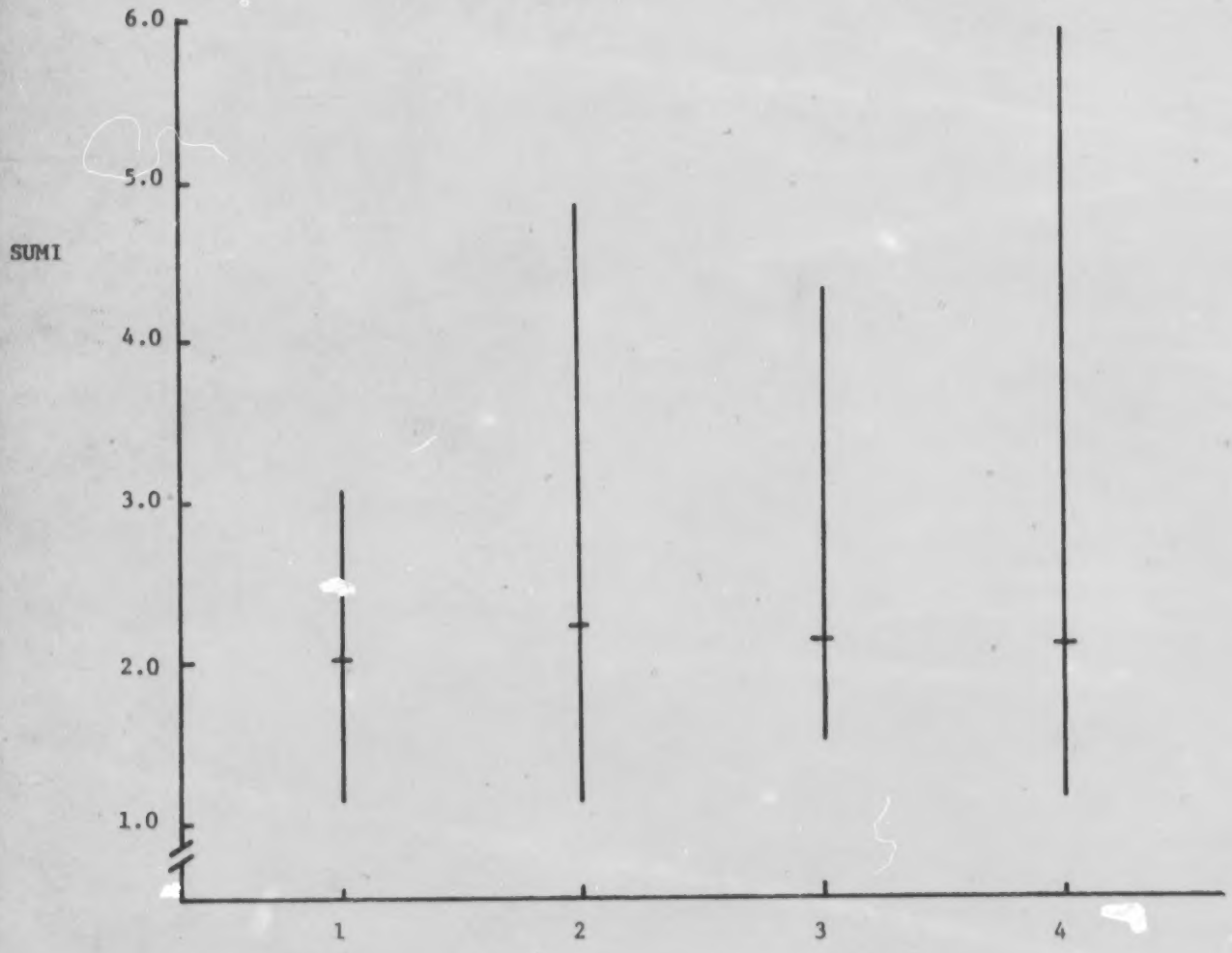
∂_c = number of column control inputs

∂_p = number of peddle control inputs

The total power shows the pilots in groups 3 and 4 made less control inputs. than groups 1 or 2 and group 3 rather clearly has the least amount of variability although not the lowest overall score. In contrast, figure 5, the H+V variable described in task 1, shows that all groups managed to control their aircraft adequately and maintained the proper rate of climb to complete the task.

Task 3

Figures 6 and 7 show data from the SUMI, and total control power variables during landing in the presence of a wind shear. Figure 6 shows groups 2 and 3 performing better than either groups 1 or 4 and having less variability indicating the aircraft was under better control for those trained in these groups. Figure 7 shows the control power. Here, group 2 shows the least overall control power input but greatest variability while group 1 shows the greatest amount of control activity (work performed by the pilot).



1006

FIGURE 1. SUMI SCORES FOR EACH GROUP (TASK 1)

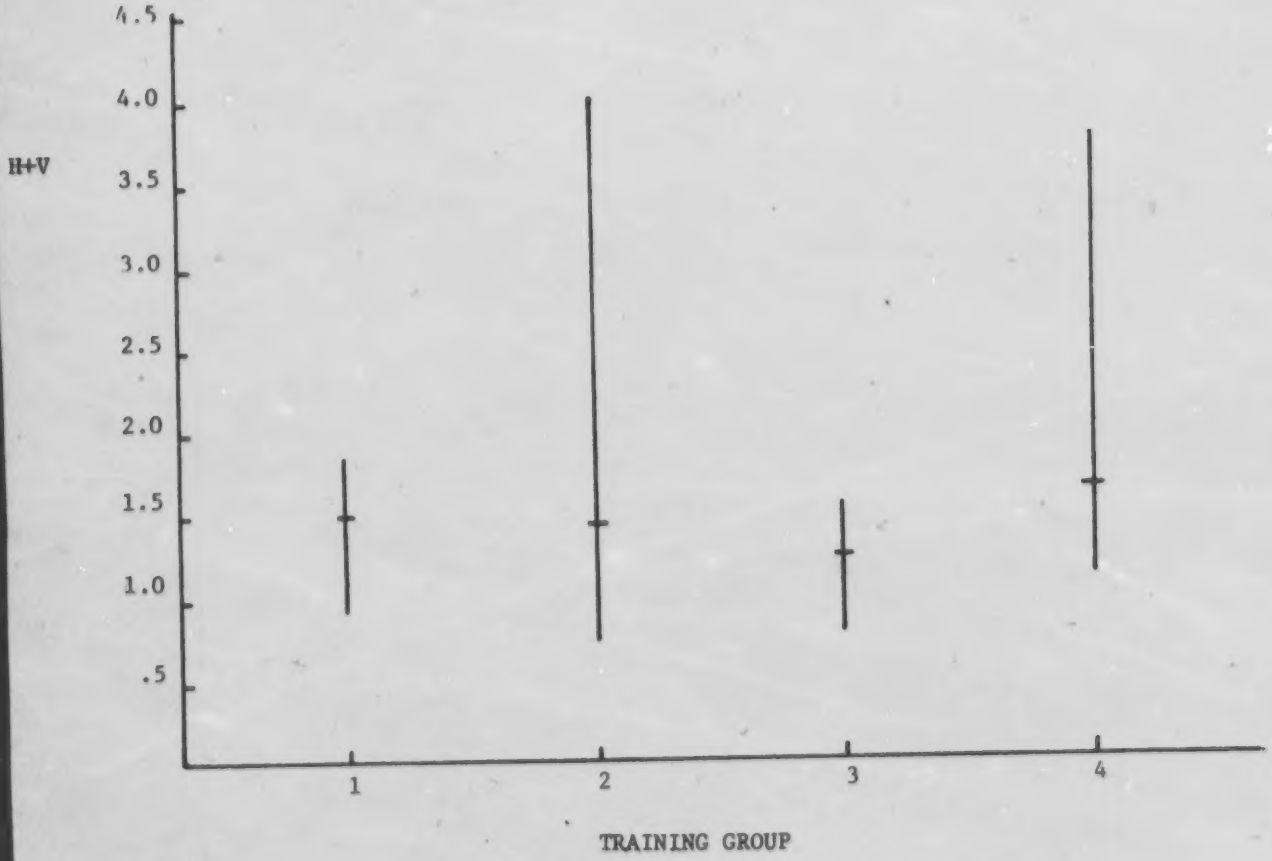


FIGURE 2 H+V SCORES FOR EACH GROUP (TASK 1)

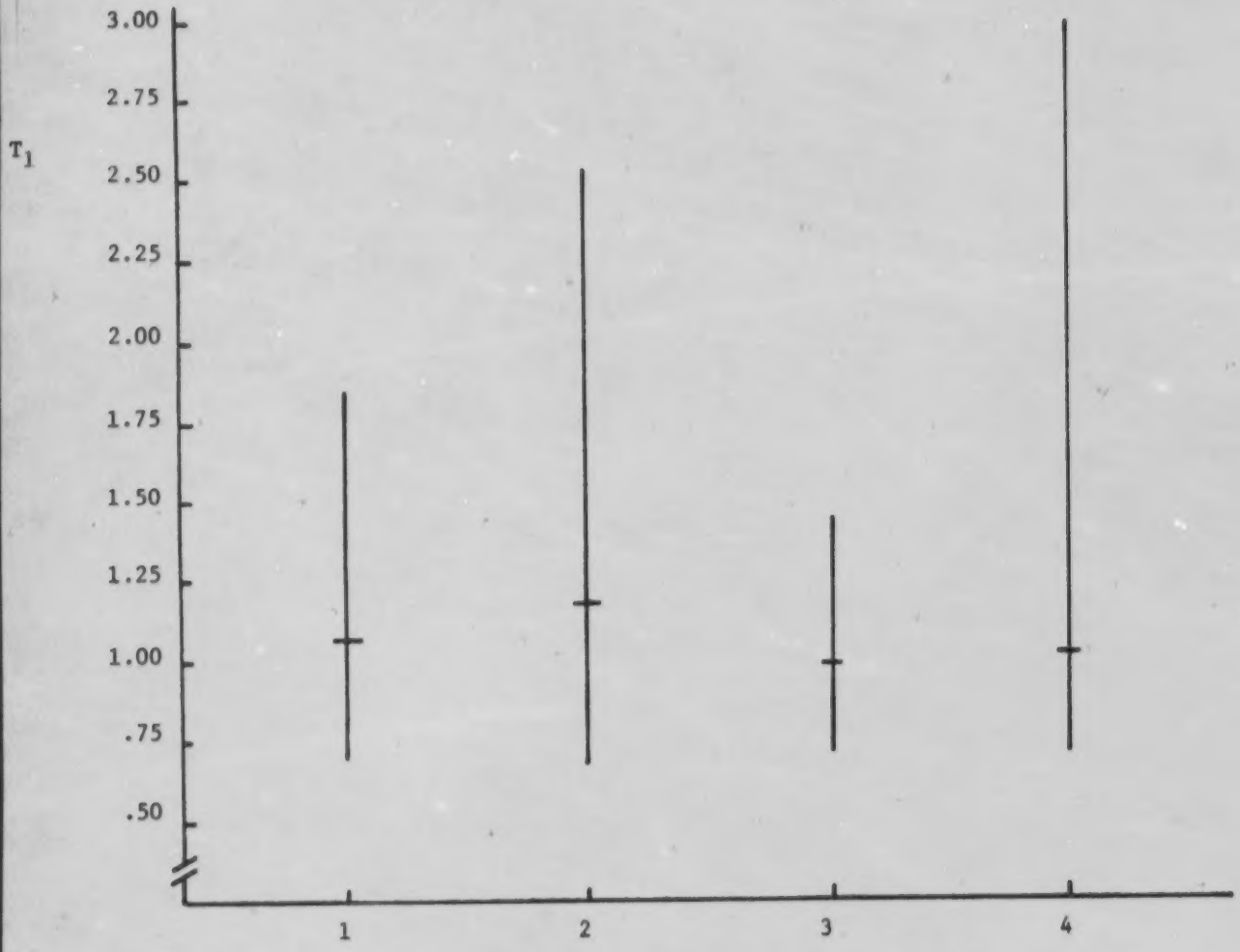


FIGURE 3 T_1 SCORES FOR EACH GROUP (TASK 1)

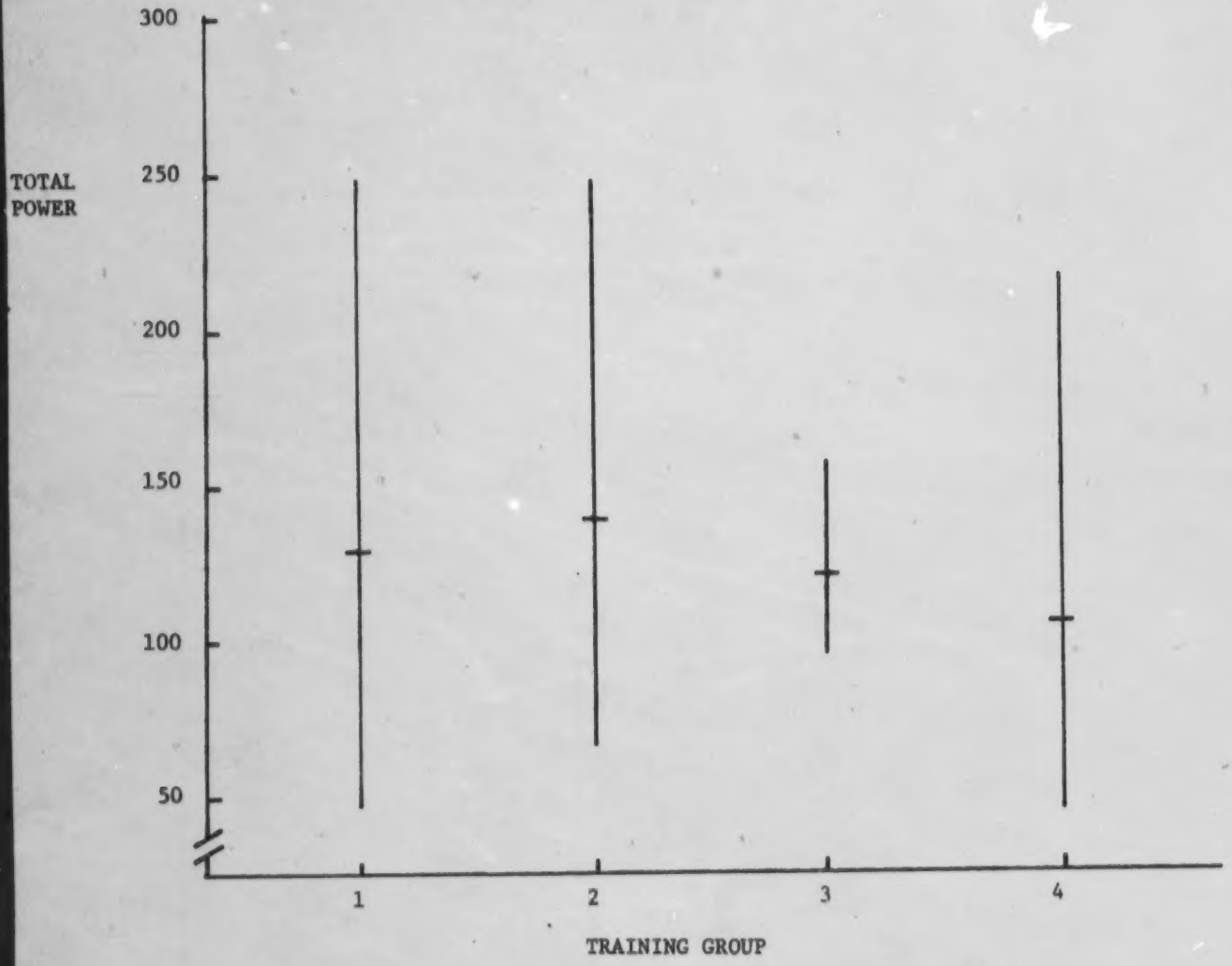


FIGURE 4 TOTAL POWER SCORES FOR EACH GROUP (TASK 2)

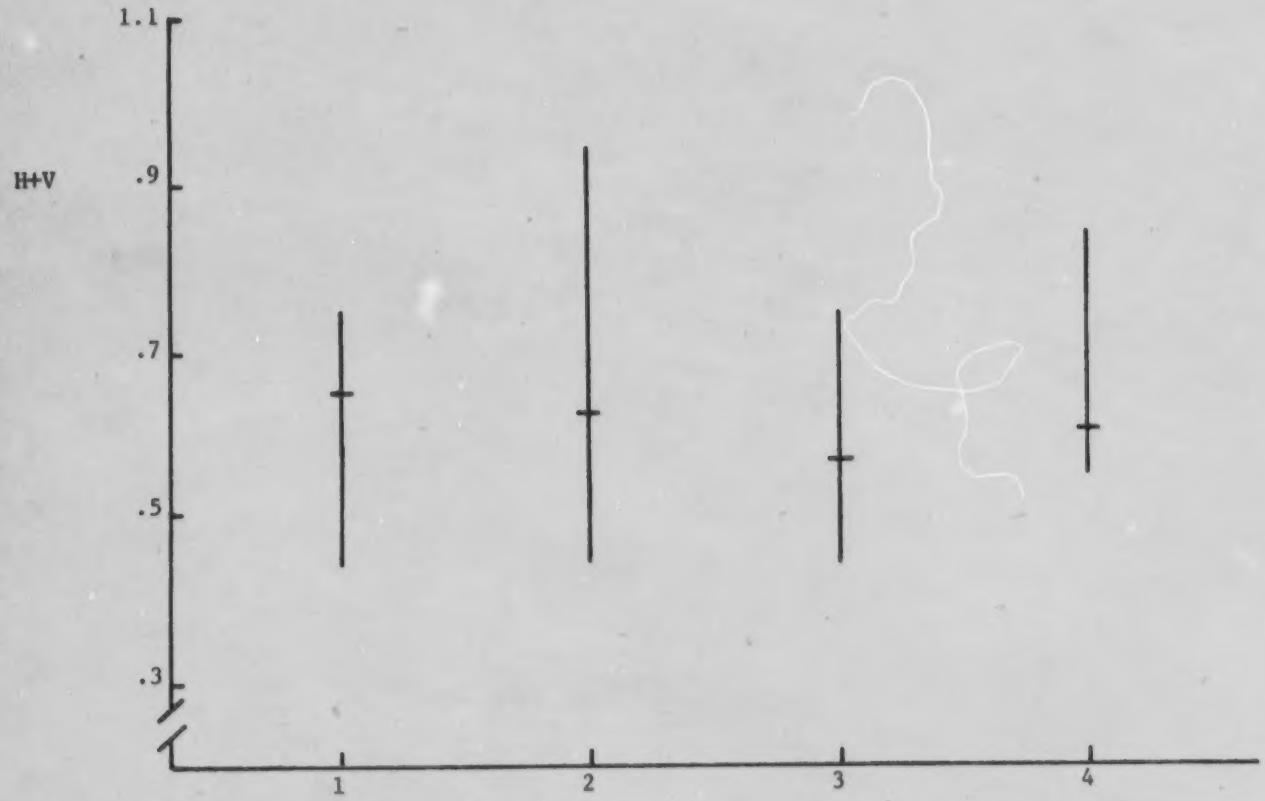


FIGURE 5 H+V SCORES FOR EACH GROUP (TASK 2)

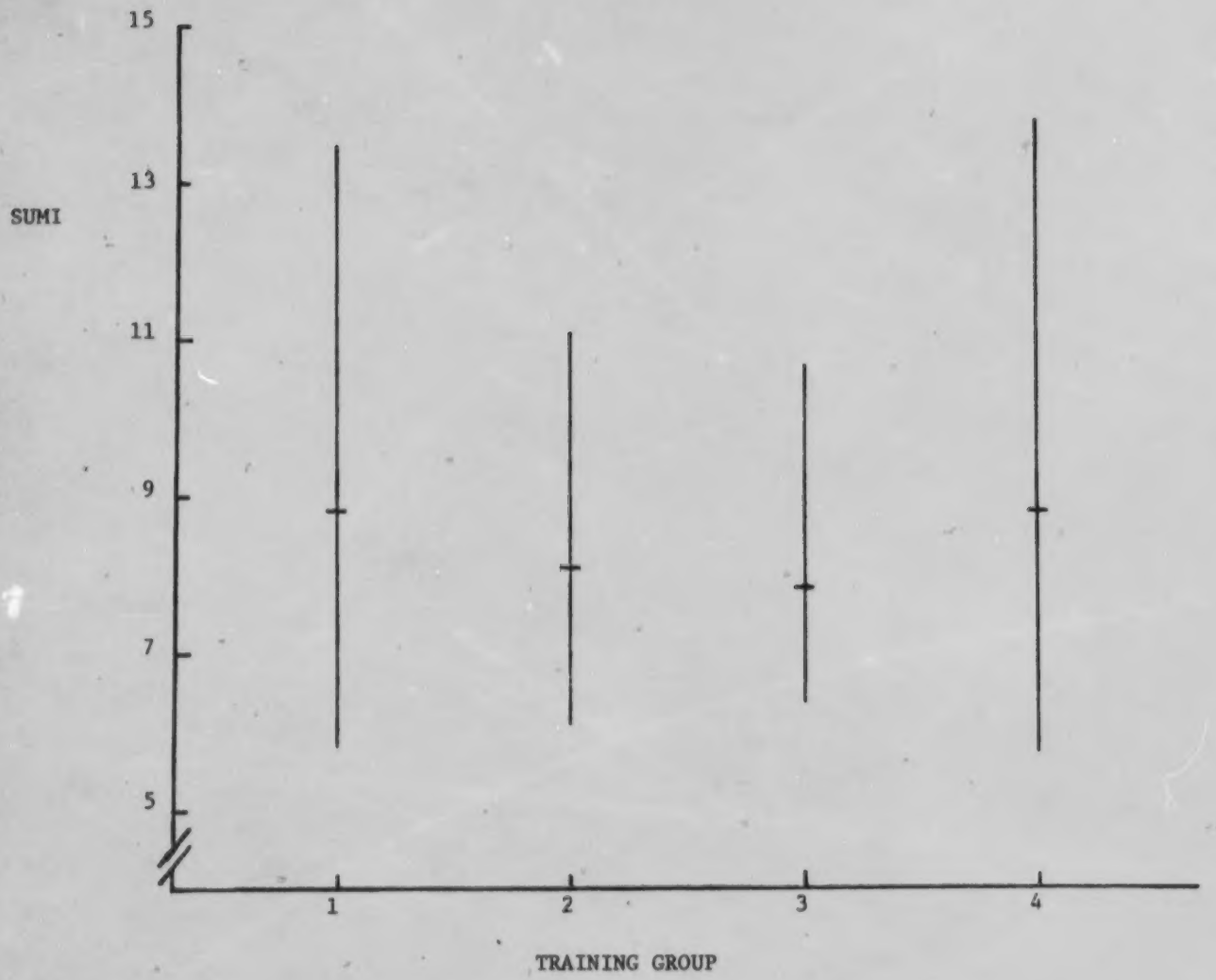


FIGURE 6 SUMI SCORES FOR EACH GROUP (TASK 3)

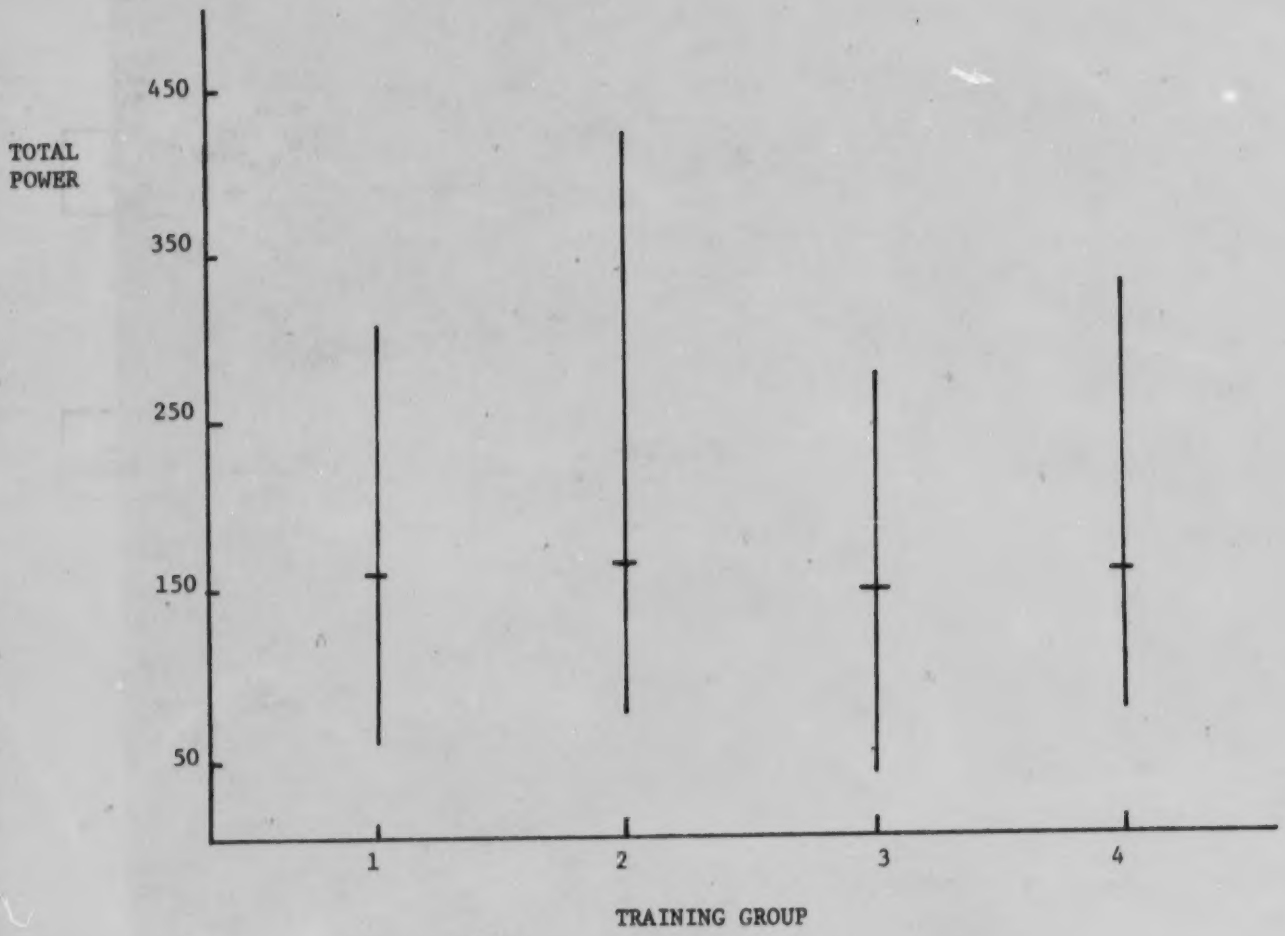
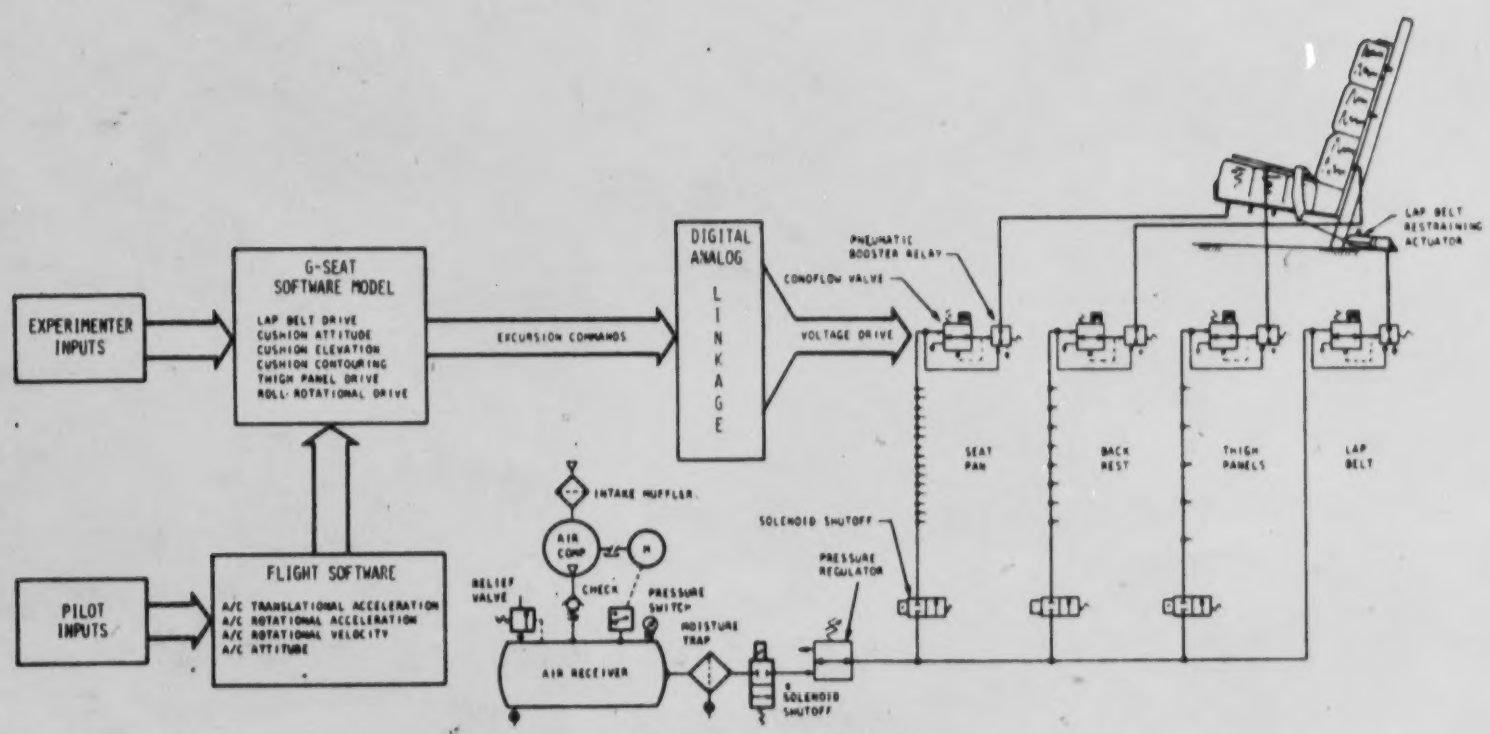


FIGURE 7 TOTAL POWER SCORES FOR EACH GROUP (TASK 3)



1013

Figure & G-SEAT SYSTEM BLOCK DIAGRAM

1014

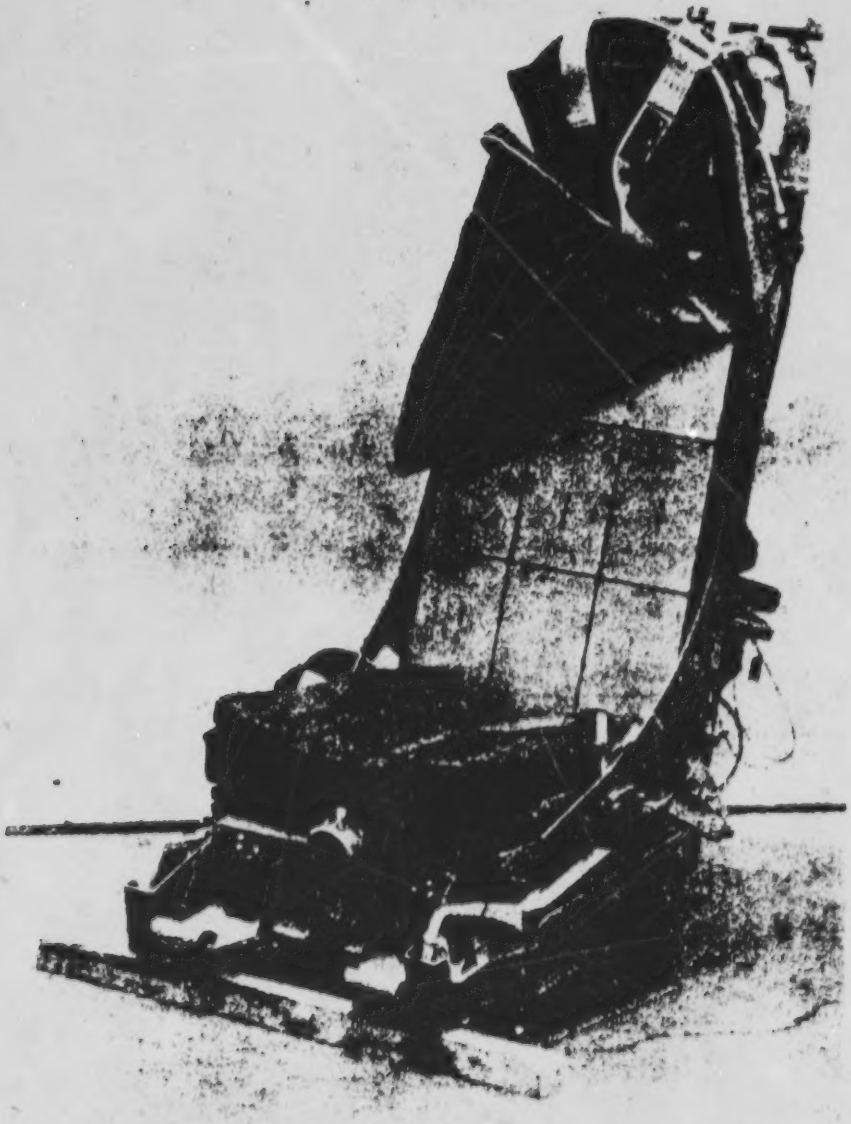


FIGURE 9 G-SEAT IN FINAL CONFIGURATION

IV. DISCUSSION

Result obtained this far appear to suggest that the g-seat does not contribute to the training effectiveness of the simulator. For a majority of the variables considered, the scatter pilot show either little difference between the g-seat and no g-seat groups, or better scores for the no g-seat groups. This was the case for all three of the tasks. Another preliminary observation of the data seems to indicate the platform motion does improve performance in a number of areas.

These conclusions are, of course, very subjective. As can be seen from the plots presented, there is a great deal of variability from task to task, from variable to variable, and even from subject to subject. Many factors influencing the results have not, as of yet, been considered. There appears to be a non-additivity of treatment effects, which is critical to the assumptions of the experimental design. These problems must wait further statistical analysis and interpretation before definitive answers can be obtained.

The reason for presenting this initial data is to provide the systems manager with at least a partial base line with which to make his decisions. The fact that the g-seat used in this study did not aid in training does not mean that future g-seat or drive systems will not enhance training effectiveness. Hardware is constantly being improved. The fact that platform motion tends to have some value for certain variables on certain tasks is not important in and by itself. The question the purchaser must answer is, are the tasks and variables investigated relevant to his particular training environment? Today we are forced to present our somewhat limited, and often conflicting results defining what cues are required in a training simulator. Ultimately the managers and users must weigh the relative cost and benefits of one type versus the other based on our results to determine what will be purchased.

REFERENCES

Parrish, R. V., Dieudonne, J. E., & Martin, D. J. Motion Software for a Synergistic Six-Degree-of-Freedom Motion Base, NASA TN-D-7350, December 1973.

Conrad, B. Operational Development of Motion Washout Techniques, NASA CR-137519, December 1973.

Kron, G. J. Advanced Simulation In Undergraduate Pilot Training: G-Seat Development, Air Force Human Resources Laboratory; AFHRL-TR-75-59(111); October 1975.

Kron, G. J. G-Seat Computer Programs and Documentation, ASUPT-74, Volume VII.

Gum, D. R. Modeling of the Human Force and Motion Sensing Mechanisms, Thesis, Ohio State University, AFHRL-TR-72-54, 1972.

DeBerg, O. H., McFarland, B. P., & Showalter, T. W. The Effect of Simulator Fidelity on Engine Failure Training in the KC-135 Aircraft. AIAA Visual and Motion Simulation Conference; April 1976.

PREDICTION OF TECHNICAL TRAINING SUCCESS FROM ASVAB AND EDUCATIONAL BACKGROUND

Lonnie D. Valentine, Jr.
Air Force Human Resources Laboratory

Since World War I, a frequent and consistent research finding has been that Blacks generally perform more poorly on test measures than do Whites. This general finding has held regardless of the test's cognitive content. It has been assumed by some that the lower test performance of Blacks does not represent their true capacity, but rather reflects social, economic, and educational deprivation. As a consequence, it has become fashionable to attack test measures as being unfair to minority members of society and, consequently, irrelevant to the accurate prediction of later performances. More recently, similar concerns about employment opportunities for women have been voiced, especially with respect to mechanical and other traditionally male jobs. However, there has been relatively little research of note with respect to sex fairness of tests.

Numerous studies have been designed to assess test fairness or to seek alternative measures which show no essential differences in the performance of various cultural subgroups. It has been noted by a number of researchers that a test can be described as biased only in the context of a later criterion event. Group difference in test performance, no matter how large, is not indicative of measurement bias when the difference is associated with a comparable difference in a criterion of concern. As early as 1953, Mary Agnes Gordon reported a bias study in which such a definition was implicit; she found that regression equations of Final Tech School grade on aptitude composites was essentially the same for Whites and Blacks and concluded that the use of the same minimum qualification scores was justified.

Other studies (e.g., Lopez, 1966; Kirkpatrick et al., 1968) have tended to substantiate the claim that Black criterion performance may be underestimated by selection procedures, while still a different group of studies (Tenopyr, 1967; Campbell et al., 1969; Guinn et al., 1970) have found that Black criterion performance tends to be overestimated by tests. Many of these studies have been subjected to methodological criticism, and numerous models for fairness in selection test use have been proposed to optimize various definitions of equitable or fair employment opportunity.

Much of the commentary appearing over the past decade with respect to test bias has missed the critical point (from a research view) that tests only provide measurement of current status of a subject with respect to a definite and limited domain. As such, they reflect group differences that may exist with respect to their domain, but they are never the cause

of the difference. It is important that research demonstrate not only overall predictive utility of selection measures but utility within various subpopulations and similarity of relationship as well. Beyond that, decisions about the way in which a valid test is used are policy matters and should provide choice of a definition of fairness based on policy considerations. For this reason, analyses presented in this paper are limited to consideration of validity and homogeneity of regression for subpopulations.

In September 1973, the Air Force discontinued use of the Armed Forces Qualification Test (AFQT) and the Airman Qualifing Examination (AQE) for non-prior service enlisted selection and initial classification in favor of the Armed Services Vocational Aptitude Battery (ASVAB). In computation of Aptitude Indexes (AIs) for the AQE, extra raw score points were awarded for completion of certain high school courses. This was based on a series of studies which had demonstrated unique predictive validity for this course information in the context of the test data.

In the conversion to ASVAB, educational points were dropped from the composites primarily because of subsequent analyses which indicated that, in operational application, their contribution to validity was relatively minor (usually enhancing validity by less than .05 correlational points), but that their inclusion increased correlation among composites by an appreciable amount.

Subsequently, Air Staff requested investigation of the ethnic fairness of educational data for classification. This request was based on an unsubstantiated allegation (which cited earlier Air Force studies) that fairer selection and classification could be achieved via educational data than from test scores.

A review of the study files from the earlier Air Force analyses indicated that:

(a) Race and sex were never included as variables in the studies.

(b) Typically, validity of the AQE's selector composite for a specialty was higher than the validity of a composite of educational variables.

(c) Educational variables contribute significantly to test variables in the prediction of training success, but that the unique contribution of these variables was less than the unique contribution of test variables.

Analyses reported here were designed to (a) investigate validity of the ASVAB and of educational data for Air Force Technical Training, (b) investigate predictive contribution of educational data to the test data,

and (c) assess homogeneity of prediction equations for subgroups defined by race and sex. Data assembled for these analyses cover Air Force accessions for Sep 73 through Oct 75, a period when ASVAB Form 3 was used for Air Force production testing. Data reported here are far from exhaustive; files developed for these analyses are currently being augmented with full ASVAB test protocols and will provide a base for other studies.

A basic working file was developed from a collation of the Air Force's Processing and Classification of Enlistees (PACE) file with Technical Training files. Frequency counts of cases in various Technical Courses were obtained; on the basis of these counts and consideration of course similarity, specialties were aggregated into 45 clusters for analysis.

Predictor variables for the analyses consisted of an AFQT score and four Air Force AIs, all derived from ASVAB-3, and a series of 41 binary variables indicating successful completion or non-completion of 41 different high school courses. Criteria were disposition from Technical Training and, for graduates, final school grade.

For each of the 45 clusters, half of the available Caucasian males were selected as an Education Index (EI) development sample; restriction to this one group was to avoid depletion of minority cases for later phases of the analyses. For each of the 45 clusters, the EI sample was divided into an upper and lower criterion group from consideration of the two criterion variables. Failure cases for whom no final course grade was available were assigned to the lower group along with those graduates with the lowest final course grades. The 42 binary high school course variables were item analyzed against this dichotomy, and the significantly positively correlated (at the .05 level or better) variables were assigned a scoring weight of +1 while those showing significant negative correlation were assigned a scoring weight of -1. The EI development samples were excluded from all succeeding analyses; thus, all validities obtained in the study represent cross validation values.

For each remaining case in each of the 45 clusters, the educational variables were scored to obtain an EI (essentially, the binary course completion variables were treated as items and scored ''Rights'' minus ''Wrongs'' to obtain a single education score).

Validities of both test measures and of the EI were computed for subsamples defined by race and by sex in each of the 45 clusters. To accomplish this, the sample was first divided into subsamples of Caucasians, Blacks, and other minority members, and validations were accomplished for these race subsamples. They were then redivided into subsamples of males and females, and the validations were accomplished for sex subsamples.

Criterion used for these validations was Final School Grade. These within subgroup validations were accomplished only for subgroups with 24 or more cases. Table 1 shows N's for subsamples on which these within group analyses were computed.

Table 2 shows means and standard deviations of the EI, the Selector AI (i.e., the ASVAB composite used for selection for the specialty), and the Final School Grade for the race defined subgroups while Table 3 shows these statistics for sex defined subgroups.

Generally, minority race means on all three variables tended to be lower than Caucasian means. With respect to the sex breakout, results were mixed; generally, female means on the Selector AI for mechanical specialties tended to be considerably lower than that of males.

Table 4 reports the multiple correlation of the EI and Selector AI with Final School Grade and their separate zero-order validities for each of the race defined subgroups. In addition, it reports F ratios for contribution of each of these two predictors to the multiple. Table 5 presents these same data for subgroups defined by sex. It should be emphasized that data presented in these two tables are the obtained values; they have <u>not</u> been subjected to correction for range restriction. Such corrections would have been improper since the Selector AI, on which correction occurred, has a rectangular distribution; for those clusters with a high selection cut score, corrections in these data would have overestimated population correlations.

In these tables, all F ratios not marked by a symbol are significant at the .01 level. Inspection of Table 4 shows that, generally, both the EI and the Selector AI are valid for predicting Final School Grade across race subsamples with the test score typically being more valid than the educational data. In addition, both kinds of data generally contribute significantly to prediction. In Table 5, similar results are also noted for subsamples based on sex.

To ascertain homogeneity of the separate race and sex prediction equations, a series of regression models was established and compared via the F statistic. The full model consisted of a series of binary subsample membership variables (for race or sex as appropriate), the EI, the Selector AI, and the product (or interactions) of these last two variables with the membership variables. Comparison of R^2 from this model with R^2 based on Selector AI and EI only tests the hypothesis that race (or sex) regressions are homogeneous. If this comparison is significant, then comparison of the full model with a model from which the binary membership variables have been excluded tests the hypothesis that regression intercepts are homogeneous, and comparison of the full model with a model from which the

interaction variables are excluded tests for homogeneity of regression slopes.

These analyses for both race and sex effects are summarized in Table 6. Generally, there are stronger effects attributable to race than to sex; i.e., in most instances, use of a separate equation for each of the three race groups adds significantly to prediction (in 37 of the 43 comparisons for H_1). Significant improvement in predictions from separate sex equations was found in 19 of 37 comparisons.

For those groups in which significant differences were found on H_1, two additional hypotheses were tested: (a) that the slopes of the regression lines were homogeneous (H_2) and (b) that the intercepts of the regression lines were homogeneous. In 10 of the 37 instances, significant race differences in slope were found with two instances of significant race differences in intercept. Eight instances (out of 19) of significant sex differences in slope were found, and seven instances of significant intercept differences.

Thus, these analyses suggest that, while predictions based on joint consideration of test and educational data have useful validity across race and sex groups, selection strategies which consider race and sex may further improve the system.

Since most such analyses of test data alone have suggested homogeneous prediction equations, it is possible that the inclusion of the educational data may be the source of the significant differences found here. These data will be further analyzed to investigate this possibility.

In summary, these data indicate that (a) both test and education background data contribute uniquely to predictions of Air Force Technical Training performance, (b) generally, the contribution of test data is larger than that of educational background, (c) these findings hold across race and sex subgroups, and (d) there was evidence that race and sex unique prediction equations, based on test and educational data, are not homogeneous.

More intensive analysis of the data base for this study will be conducted. At the present time, item response data are being added to the files; this will allow generation of all subtest and raw composite scores. Later investigations will examine appropriateness of composites as presently constituted, seek more valid composites, consider the number of composites needed, and will examine fairness of these with respect to both race and sex. These analyses will also examine homogeneity of prediction equations for test and educational data separately.

Table 1. Within Subgroup Sample Sizes

Group Nr.	Job Area	Cauc N	Black N	Other Min. N	Male N	Female N
01	Intelligence (20X30)	245	43	-	235	55
02	Morse Systems Operator (20731)	229	78	-	233	78
03	Audiovisual (23X30)	171	43	-	183	31
04	Weather (25X3X)	317	55	-	278	96
05	Command Control Systems Operator (27X3X)	664	230	-	790	115
06	Communications Operations (29130)	369	195	-	409	158
07	Radio Operations (29333)	258	43	-	248	59
08	Communications-Electronics Systems (30X3X)	1849	181	53	1740	343
09	Missile Electronic Maintenance (31X3X)	544	53	-	517	95
10	Avionics Systems (32X3X)	2163	244	57	2013	450
11	Training Devices (34X3X)	170	-	-	158	-
12	Wire Communications Systems Maintenance (361/3X0)	226	66	-	303	-
13	Wire Communications Systems Maintenance (362X0)	224	69	-	287	-
14	Intricate Equipment Maintenance (40X3X)	75	24	-	101	-

Table 1 (continued)

Group Nr.	Job Area	Cauc N	Black N	Other Min. N	Male N	Female N
15	Aircraft Accessory Maintenance (42X3X)	1598	1041	98	2187	550
16	Aircraft Maintenance (43130)	193	-	-	177	44
17	Aircraft Maintenance (43131)	4559	1073	104	4468	1268
18	Aircraft Engineer (4323X)	1356	363	44	1431	332
19	Missile Maintenance (44X3X)	241	52	-	259	36
20	Munitions & Weapons Maintenance (46130)	832	162	-	1008	-
21	Munitions & Weapons Maintenance (46230)	912	154	-	1084	-
22	Munitions & Weapons Maintenance (46330)	194	-	-	208	-
23	Vehicle Maintenance (47X3X)	251	28	-	282	-
24	Computer Systems (51X3X)	251	-	-	183	86
25	Metal Working (53X3X)	653	160	-	659	168
26	Mechanical/Electrical (54X3X)	831	297	-	970	181
27	Structural/Pavements (55X3X)	505	75	-	471	119
28	Sanitation (56330)	215	36	-	251	-

Table 1 (continued)

Group Nr.	Job Area	Cauc N	Black N	Other Min. N	Male N	Female N
29	Fire Protection (57130)	507	188	-	709	-
30	Fabric and Rubber Products (58X30)	178	42	-	194	29
31	Transportation (60X3X)	1106	400	40	1346	200
32	Food Service (62X3X)	256	136	-	284	117
33	Fuel Services (63130)	367	265	-	644	-
34	Inventory Management (64530)	1199	587	83	1312	556
35	Material Facilities (64730)	481	360	-	541	317
36	Accounting and Finance, and Auditing (67X3X)	439	100	-	372	179
37	Administration (70X3X)	1503	1078	56	1716	921
38	Personnel (73230)	453	180	-	463	185
39	Security Police (81130)	2172	1222	44	3438	-
40	Law Enforcement and Corrections (81230)	1078	256	-	900	448
41	Medical (90010)	934	404	28	912	454
42	Medical (90X3X)	1385	470	48	1283	620

Table 1 (continued)

Group Nr.	Job Area	Cau N	Black N	Other Min. N	Male N	Female N
43	Medical (91X3X)	249	48	-	251	49
44	Aircrew Protection (92230)	332	63	-	339	63
45	Dental (98X3X)	241	68	-	212	108

Table 2. Within Ethnic Group Means and Standard Deviations

Group Nr.	Sample	Educ Index $\bar{X}$	Educ Index S	Selector AI $\bar{X}$	Selector AI S	Final Course Grade $\bar{X}$	Final Course Grade S
01	Cauc	1.58	1.67	82.57	9.86	85.71	6.26
	Black	1.44	1.59	80.34	7.42	81.07	5.36
	Other	-	-	-	-	-	-
02	Cauc	-.27	.52	75.61	10.87	-	-
	Black	-.23	.48	70.38	9.60	-	-
	Other	-	-	-	-	-	-
03	Cauc	4.09	1.65	72.34	10.78	82.58	6.90
	Black	3.65	1.67	68.14	9.28	76.93	8.00
	Other	-	-	-	-	-	-
04	Cauc	3.46	1.80	88.19	5.56	89.66	4.92
	Black	3.73	1.66	85.00	4.67	84.64	4.37
	Other	-	-	-	-	-	-
05	Cauc	2.53	1.67	74.89	12.14	86.82	5.36
	Black	2.80	1.64	69.89	12.29	84.36	5.21
	Other	-	-	-	-	-	-
06	Cauc	2.96	1.69	72.09	10.10	85.64	7.18
	Black	2.73	1.77	68.62	8.45	83.13	7.66
	Other	-	-	-	-	-	-
07	Cauc	.36	.63	71.43	10.93	-	-
	Black	.30	.67	69.07	10.07	-	-
	Other	-	-	-	-	-	-
08	Cauc	2.88	2.69	85.01	7.29	85.09	6.01
	Black	2.66	2.70	82.35	6.98	81.83	5.34
	Other	4.04	2.58	82.36	8.04	84.60	6.42
09	Cauc	3.53	2.02	84.60	7.68	86.01	6.14
	Black	4.11	1.94	82.83	5.62	83.00	6.34
	Other	-	-	-	-	-	-
10	Cauc	6.73	2.83	85.26	7.17	84.35	6.38
	Black	6.62	2.94	83.32	6.96	81.12	6.44
	Other	7.07	3.13	85.70	6.91	85.28	5.95

Table 2 (continued)

Group Nr.	Sample	Educ Index $\bar{X}$	Educ Index S	Selector AI $\bar{X}$	Selector AI S	Final Course Grade $\bar{X}$	Final Course Grade S
11	Cauc	3.34	2.09	85.71	6.56	84.94	5.22
	Black	-	-	-	-	-	-
	Other	-	-	-	-	-	-
12	Cauc	1.47	.93	60.95	16.24	78.69	7.18
	Black	1.50	.82	46.21	10.59	75.58	5.92
	Other	-	-	-	-	-	-
13	Cauc	.89	.67	67.63	11.98	78.55	6.46
	Black	.99	.77	65.80	8.58	75.65	5.98
	Other	-	-	-	-	-	-
14	Cauc	.84	1.24	74.13	10.75	82.60	6.78
	Black	.33	1.11	64.17	4.93	77.83	4.05
	Other	-	-	-	-	-	-
15	Cauc	3.90	2.21	48.87	21.78	83.13	7.56
	Black	3.75	2.15	36.58	15.50	80.41	6.62
	Other	4.28	2.61	37.09	16.75	81.62	7.39
16	Cauc	2.41	1.79	68.60	21.73	81.48	7.32
	Black	-	-	-	-	-	-
	Other	-	-	-	-	-	-
17	Cauc	1.35	2.02	62.06	20.62	84.42	7.78
	Black	1.31	2.04	51.23	17.12	79.19	7.92
	Other	1.62	2.36	56.97	17.14	82.57	7.26
18	Cauc	3.89	1.89	64.32	17.66	85.31	7.41
	Black	3.81	1.82	51.60	12.08	80.02	7.00
	Other	4.11	2.04	53.64	14.82	84.27	7.55
19	Cauc	1.44	.98	64.77	12.61	84.68	6.38
	Black	1.54	1.05	56.54	7.88	80.21	5.52
	Other	-	-	-	-	-	-
20	Cauc	2.45	1.68	72.35	12.65	89.28	4.99
	Black	2.81	1.59	60.59	14.14	86.39	5.29
	Other	-	-	-	-	-	-

Table 2 (continued)

Group Nr.	Sample	Educ Index $\bar{X}$	Educ Index S	Selector AI $\bar{X}$	Selector AI S	Final Course Grade $\bar{X}$	Final Course Grade S
21	Cauc	2.20	1.81	72.98	12.56	89.52	5.05
	Black	2.26	1.79	63.08	12.77	86.19	5.24
	Other	-	-	-	-	-	-
22	Cauc	4.48	2.03	84.59	8.39	90.77	4.42
	Black	-	-	-	-	-	-
	Other	-	-	-	-	-	-
23	Cauc	.90	1.55	73.80	16.48	79.23	7.33
	Black	.64	1.39	55.00	11.95	72.00	6.51
	Other	-	-	-	-	-	-
24	Cauc	.09	.29	86.49	10.47	84.30	7.01
	Black	-	-	-	-	-	-
	Other	-	-	-	-	-	-
25	Cauc	3.00	1.83	63.12	17.13	84.54	6.99
	Black	2.92	1.80	53.16	14.25	79.51	6.86
	Other	-	-	-	-	-	-
26	Cauc	1.85	1.73	57.18	20.73	82.23	7.38
	Black	1.91	1.63	41.90	16.52	78.64	6.96
	Other	-	-	-	-	-	-
27	Cauc	2.51	1.52	61.24	17.46	82.45	7.04
	Black	1.95	1.43	47.53	15.52	76.39	7.01
	Other	-	-	-	-	-	-
28	Cauc	2.20	1.65	59.51	17.10	80.92	6.80
	Black	2.83	1.96	47.78	10.03	75.61	7.20
	Other	-	-	-	-	-	-
29	Cauc	2.67	1.61	66.68	14.64	87.05	5.23
	Black	2.71	1.49	59.65	13.15	82.52	5.46
	Other	-	-	-	-	-	-
30	Cauc	2.39	1.91	54.61	15.74	85.29	5.43
	Black	2.21	1.55	44.88	8.13	81.88	4.74
	Other	-	-	-	-	-	-

Table 2 (continued)

Group Nr.	Sample	Educ Index $\bar{X}$	Educ Index S	Selector AI $\bar{X}$	Selector AI S	Final Course Grade $\bar{X}$	Final Course Grade S
31	Cauc	2.41	1.79	53.92	18.22	83.11	6.58
	Black	2.41	1.58	51.10	15.44	79.08	5.94
	Other	3.38	2.06	51.38	14.70	82.75	5.83
32	Cauc	.51	.60	60.37	13.70	87.57	6.38
	Black	.44	.66	57.50	11.66	84.68	7.69
	Other	-	-	-	-	-	-
33	Cauc	.83	.88	63.04	15.40	91.73	5.26
	Black	.86	.87	56.77	11.98	87.45	6.10
	Other	-	-	-	-	-	-
34	Cauc	3.31	1.92	68.58	14.35	84.75	7.14
	Black	3.18	1.86	62.81	16.02	81.92	6.86
	Other	4.61	2.25	67.95	16.30	83.61	7.26
35	Cauc	1.04	1.08	63.07	14.32	83.53	7.38
	Black	1.08	1.09	57.40	12.10	80.96	7.16
	Other	-	-	-	-	-	-
36	Cauc	4.46	2.18	87.57	5.88	80.84	7.79
	Black	4.15	2.15	86.90	5.74	76.76	7.68
	Other	-	-	-	-	-	-
37	Cauc	2.86	1.81	63.67	14.76	84.55	6.57
	Black	2.86	1.71	57.29	14.14	81.69	6.34
	Other	4.27	2.17	59.46	12.98	85.43	7.12
38	Cauc	2.70	2.09	74.04	11.91	87.73	5.56
	Black	2.12	1.89	70.06	11.14	83.45	5.46
	Other	-	-	-	-	-	-
39	Cauc	8.65	2.37	62.37	14.63	86.36	6.67
	Black	8.54	2.25	56.61	12.87	81.76	7.43
	Other	8.98	2.92	61.25	13.70	83.66	8.24
40	Cauc	2.82	1.83	68.69	12.96	83.61	5.90
	Black	3.09	1.82	64.65	12.33	79.82	6.66
	Other	-	-	-	-	-	-

Table 2 (continued)

Group Nr.	Sample	Educ Index $\bar{X}$	Educ Index S	Selector AI $\bar{X}$	Selector AI S	Final Course Grade $\bar{X}$	Final Course Grade S
41	Cauc	4.05	2.19	75.20	12.05	83.79	7.66
	Black	3.76	2.00	70.80	10.38	76.83	7.84
	Other	5.61	2.43	73.75	13.34	83.29	8.05
42	Cauc	4.02	2.05	76.64	11.04	82.44	6.98
	Black	3.65	2.08	72.28	10.32	76.27	6.97
	Other	5.29	2.25	78.65	11.12	81.88	7.81
43	Cauc	2.36	1.33	76.08	10.99	80.96	6.17
	Black	2.27	1.22	71.98	11.31	78.69	6.33
	Other	-	-	-	-	-	-
44	Cauc	-1.09	.58	63.86	15.10	82.50	6.41
	Black	-1.10	.61	58.57	13.87	79.22	6.41
	Other	-	-	-	-	-	-
45	Cauc	1.46	1.51	73.78	10.36	82.95	7.78
	Black	1.29	1.38	70.51	9.08	79.00	6.83
	Other	-	-	-	-	-	-

Table 3. Within Sex Means and Standard Deviations

Group Nr.	Sample	Educ Index $\bar{X}$	Educ Index S	Selector AI $\bar{X}$	Selector AI S	Final Course Grade $\bar{X}$	Final Course Grade S
01	Male	1.55	1.69	83.28	8.46	85.50	6.23
	Female	1.62	1.50	78.18	12.52	83.20	6.53
02	Male	-.32	.55	73.61	10.89	-	-
	Female	-.08	.27	76.09	10.21	-	-
03	Male	4.02	1.65	71.97	10.63	81.36	7.58
	Female	3.94	1.72	68.71	10.16	82.00	6.90
04	Male	3.70	1.82	87.75	5.45	89.18	5.24
	Female	2.92	1.51	87.50	5.82	88.12	4.77
05	Male	2.68	1.68	74.03	12.41	86.19	5.40
	Female	2.07	1.51	70.26	11.88	85.94	5.64
06	Male	2.94	1.76	70.78	9.36	84.75	7.51
	Female	2.75	1.64	71.20	10.68	84.78	7.35
07	Male	.33	.65	70.50	10.71	-	-
	Female	.46	.56	73.14	11.04	-	-
08	Male	2.77	2.66	85.20	7.25	84.95	6.03
	Female	3.46	2.76	82.22	7.27	83.97	5.99
09	Male	3.60	2.02	85.17	7.45	85.87	6.25
	Female	3.40	1.94	80.00	6.77	85.06	6.01
10	Male	6.94	2.83	85.62	7.11	84.07	6.42
	Female	5.76	2.74	82.63	6.91	83.96	6.59
11	Male	3.52	2.11	86.01	6.48	84.55	5.29
	Female	-	-	-	-	-	-
12	Male	1.47	.91	57.66	16.25	78.02	7.04
	Female	-	-	-	-	-	-
13	Male	.93	.70	67.18	11.32	77.63	6.42
	Female	-	-	-	-	-	-
14	Male	.72	1.20	71.29	10.43	81.13	6.38
	Female	-	-	-	-	-	-

Table 3 (continued)

Group Nr.	Sample	Educ Index $\bar{X}$	Educ Index S	Selector AI $\bar{X}$	Selector AI S	Final Course Grade $\bar{X}$	Final Course Grade S
15	Male	4.13	2.16	47.53	19.93	82.45	7.29
	Female	2.77	2.05	28.86	14.36	80.44	7.26
16	Male	2.35	1.80	73.64	15.70	81.54	7.43
	Female	2.32	1.53	37.61	19.11	77.27	7.38
17	Male	1.54	2.07	66.17	16.46	83.99	7.99
	Female	.68	1.72	38.00	17.59	81.34	7.98
18	Male	3.97	1.88	64.19	17.59	84.42	7.74
	Female	3.52	1.85	49.58	10.19	83.26	7.07
19	Male	1.35	.97	64.56	12.52	83.81	6.39
	Female	2.28	.80	54.31	4.27	84.28	6.89
20	Male	2.51	1.67	70.35	13.59	88.82	5.15
	Female	-	-	-	-	-	-
21	Male	2.20	1.81	71.49	13.07	89.02	5.24
	Female	-	-	-	-	-	-
22	Male	4.44	2.03	84.11	8.47	90.45	4.57
	Female	-	-	-	-	-	-
23	Male	.87	1.53	71.76	17.06	78.51	7.53
	Female	-	-	-	-	-	-
24	Male	.10	.30	86.72	10.64	83.98	7.20
	Female	.08	.27	84.88	10.37	84.99	6.75
25	Male	2.97	1.86	64.54	16.21	83.46	7.20
	Female	3.10	1.71	47.38	13.59	83.83	7.48
26	Male	1.86	1.70	57.43	18.78	81.56	7.35
	Female	2.02	1.78	29.12	13.71	79.62	7.57
27	Male	2.72	1.46	63.13	16.89	82.00	7.22
	Female	1.26	1.14	44.50	12.69	80.15	7.52
28	Male	2.29	1.71	57.83	16.79	80.16	7.11
	Female	-	-	-	-	-	-

Table 3 (continued)

Group Nr.	Sample	Educ Index $\bar{X}$	Educ Index S	Selector AI $\bar{X}$	Selector AI S	Final Course Grade $\bar{X}$	Final Course Grade S
29	Male	2.68	1.58	64.69	14.51	85.81	5.66
	Female	-	-	-	-	-	-
30	Male	2.50	1.86	53.12	15.47	84.59	5.46
	Female	1.59	1.65	50.00	11.22	84.86	5.61
31	Male	2.39	1.74	51.13	16.98	81.87	6.61
	Female	2.72	1.76	66.58	14.87	83.34	6.72
32	Male	.49	.63	60.56	13.60	85.40	7.35
	Female	.49	.61	56.62	11.85	89.20	5.30
33	Male	.85	.87	60.40	14.45	89.88	6.04
	Female	-	-	-	-	-	-
34	Male	3.44	2.00	65.24	15.55	83.78	7.19
	Female	3.05	1.77	70.23	13.77	83.87	7.17
35	Male	1.06	1.10	60.98	13.54	82.27	6.98
	Female	1.10	1.09	59.91	13.93	82.76	8.06
36	Male	4.73	2.18	87.50	5.84	80.43	7.89
	Female	3.78	2.02	87.40	5.85	79.21	7.95
37	Male	3.05	1.83	59.24	14.53	82.98	6.58
	Female	2.60	1.67	64.20	14.78	84.20	6.69
38	Male	2.66	2.13	72.43	11.65	86.22	5.96
	Female	2.35	1.93	73.97	12.13	87.14	5.67
39	Male	8.62	2.34	60.31	14.29	84.69	7.31
	Female	-	-	-	-	-	-
40	Male	3.04	1.88	68.66	13.24	83.62	6.19
	Female	2.52	1.66	66.29	12.19	81.36	6.02
41	Male	4.03	2.17	74.40	11.35	81.33	8.36
	Female	3.93	2.12	72.82	12.53	82.52	8.27
42	Male	4.11	2.09	75.97	11.24	80.87	7.49
	Female	3.65	2.02	74.88	10.57	80.97	7.48

Table 3 (continued)

Group Nr.	Sample	Educ Index X̄	Educ Index S	Selector AI X̄	Selector AI S	Final Course Grade X̄	Final Course Grade S
43	Male	2.34	1.35	75.26	11.06	80.31	6.21
	Female	2.41	1.14	76.43	11.61	81.94	6.30
44	Male	-1.09	.57	63.32	15.44	82.37	6.48
	Female	-1.11	.67	61.43	12.17	79.75	6.08
45	Male	1.58	1.51	73.11	10.35	81.75	7.83
	Female	1.16	1.44	72.04	10.14	82.46	7.51

Table 4. Within Ethnic Group Validities and Contributions to Prediction

Group Nr.	Sample	R	Validity of: Educ Index	Validity of: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
01	Cauc	.48	.40	.38	24.46	20.10
	Black	.46	.24	.43	1.01#	7.57
	Other	-	-	-	-	-
03	Cauc	.43	.38	.31	19.49	8.57
	Black	.44	.40	.28	5.61*	1.73#
	Other	-	-	-	-	-
04	Cauc	.35	.30	.23	25.30	11.85
	Black	.36	.22	.28	2.25#	4.01#
	Other	-	-	-	-	-
05	Cauc	.35	.23	.33	7.08	49.55
	Black	.30	.20	.27	3.57#	12.02
	Other	-	-	-	-	-
06	Cauc	.40	.29	.36	14.03	34.98
	Black	.28	.18	.23	5.16*	9.24
	Other	-	-	-	-	-
08	Cauc	.48	.29	.44	87.98	359.94
	Black	.40	.24	.34	9.31	20.55
	Other	.05	-.04	.03	.07#	.06#
09	Cauc	.49	.36	.45	29.48	83.91
	Black	.41	.06	.40	8.13	9.44
	Other	-	-	-	-	-
10	Cauc	.36	.27	.32	70.81	141.49
	Black	.41	.31	.35	12.45	20.53
	Other	.29	.09	.29	.00#	4.40#
11	Cauc	.40	.35	.32	11.67	7.45
	Black	-	-	-	-	-
	Other	-	-	-	-	-
12	Cauc	.40	.30	.26	24.99	18.94
	Black	.25	.05	.25	.12#	4.03*
	Other	-	-	-	-	-

Table 4 (continued)

Group Nr.	Sample	R	Validity of: Educ Index	Validity of: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
13	Cauc	.34	.26	.28	9.43	12.48
	Black	.08	.08	.00	.43#	.00#
	Other	-	-	-	-	-
14	Cauc	.54	.44	.47	7.08	10.44
	Black	.16	.15	-.01	.51#	.04#
	Other	-	-	-	-	-
15	Cauc	.44	.29	.43	25.90	226.06
	Black	.29	.21	.23	34.66	46.13
	Other	.36	.16	.34	1.47#	11.17
16	Cauc	.46	.30	.39	13.51	27.80
	Black	-	-	-	-	-
	Other	-	-	-	-	-
17	Cauc	.37	.26	.34	143.80	372.44
	Black	.20	.18	.12	29.63	11.28
	Other	.27	.15	.24	1.69#	5.42*
18	Cauc	.47	.35	.40	113.76	168.40
	Black	.33	.22	.29	11.41	25.34
	Other	.53	.28	.52	.14#	11.09
19	Cauc	.35	.24	.24	17.40	16.48
	Black	.24	.06	.23	.17#	2.62#
	Other	-	-	-	-	-
20	Cauc	.40	.27	.32	61.27	87.64
	Black	.28	.11	.25	2.16#	10.77
	Other	-	-	-	-	-
21	Cauc	.33	.25	.26	44.23	49.55
	Black	.11	.10	.05	1.47#	.33#
	Other	-	-	-	-	-
22	Cauc	.54	.47	.46	21.69	19.84
	Black	-	-	-	-	-
	Other	-	-	-	-	-

Table 4 (continued)

Group Nr	Sample	R	Validity of: Educ Index	Validity of: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
23	Cauc	.54	.27	.51	11.83	77.92
	Black	.13	.06	.13	.00#	.32#
	Other	-	-	-	-	-
24	Cauc	.27	.14	.24	3.90*	13.67
	Black	-	-	-	-	-
	Other	-	-	-	-	-
25	Cauc	.33	.24	.26	30.34	36.17
	Black	.21	.20	-.08	6.19*	.69#
	Other	-	-	-	-	-
26	Cauc	.43	.22	.40	27.40	139.49
	Black	.26	.15	.22	5.24*	13.61
	Other	-	-	-	-	-
27	Cauc	.30	.15	.29	.91#	36.49
	Black	.24	.05	.24	.01#	3.96#
	Other	-	-	-	-	-
28	Cauc	.54	.44	.44	31.39	31.42
	Black	.35	.33	.22	2.81#	.44#
	Other	-	-	-	-	-
29	Cauc	.27	.24	.19	17.49	6.63*
	Black	.19	.15	.12	3.83#	2.18#
	Other	-	-	-	-	-
30	Cauc	.46	.37	.39	11.91	16.35
	Black	.24	-.13	.21	.63#	1.65#
	Other	-	-	-	-	-
31	Cauc	.35	.32	.24	81.54	28.79
	Black	.23	.22	.00	20.49	1.48#
	Other	.10	.09	-.05	.28#	.03#
32	Cauc	.07	.07	.03	1.04#	.05#
	Black	.16	.09	-.16	.50#	2.66#
	Other	-	-	-	-	-

Table 4 (continued)

Group Nr.	Sample	R	Validity of: Educ Index	Validity of: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
33	Cauc	.29	.17	.28	2.08#	21.35
	Black	.19	.17	.12	5.33*	1.50#
	Other	-	-	-	-	-
34	Cauc	.30	.28	.15	90.55	12.72
	Black	.26	.26	.03	40.71	.11#
	Other	.21	.19	.10	2.64#	.48#
35	Cauc	.36	.21	.34	7.85	46.29
	Black	.16	.15	.07	7.13	.33#
	Other	-	-	-	-	-
36	Cauc	.27	.27	.04	32.34	.00#
	Black	.23	.22	-.06	4.97*	.51#
	Other	-	-	-	-	-
37	Cauc	.29	.23	.22	55.12	48.85
	Black	.24	.23	.08	56.23	2.75#
	Other	.19	.18	.10	1.37#	.24#
38	Cauc	.39	.33	.26	43.96	23.37
	Black	.31	.31	.07	17.73	.15#
	Other	-	-	-	-	-
39	Cauc	.30	.27	.21	104.26	39.33
	Black	.20	.20	.05	45.39	.14#
	Other	.52	.50	.28	10.77	1.22#
40	Cauc	.43	.32	.39	44.04	112.14
	Black	.38	.33	.29	17.28	3.91*
	Other	-	-	-	-	-
41	Cauc	.45	.38	.38	71.65	67.20
	Black	.22	.19	.12	13.64	3.18#
	Other	.26	.12	.25	.16#	1.40#
42	Cauc	.44	.35	.37	96.84	116.75
	Black	.32	.26	.26	20.74	18.93
	Other	.45	.19	.45	.19#	9.17

Table 4 (continued)

Group Nr.	Sample	R	Validity of: Educ Index	Validity of: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
43	Cauc	.43	.31	.37	14.50	26.85
	Black	.36	.30	.27	2.76#	1.98#
	Other	-	-	-	-	-
44	Cauc	.23	.21	.11	14.41	3.73#
	Black	.12	.08	-.11	.23#	.49#
	Other	-	-	-	-	-
45	Cauc	.50	.43	.40	27.59	20.06
	Black	.56	.37	.51	4.91*	15.96
	Other	-	-	-	-	-

\# Not significant.
* Significant at the .05 level. All others significant at the .01 level.

Table 5. Within Sex Validities and Contributions to Prediction

Group Nr.	Sample	R	Validity for: Educ Index	Validity for: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
01	Male	.45	.38	.37	19.43	17.14
	Females	.50	.39	.43	4.75*	7.04*
03	Male	.45	.39	.31	23.51	10.56
	Female	.53	.43	.48	2.03#	3.82#
04	Male	.40	.26	.34	13.58	30.51
	Female	.19	.17	.11	2.22#	.44#
05	Male	.36	.21	.35	5.74*	81.33
	Female	.33	.29	.29	2.99#	3.47#
06	Male	.39	.25	.35	12.78	40.67
	Female	.38	.29	.32	7.78	11.15
08	Male	.47	.29	.43	79.42	311.15
	Female	.50	.29	.47	14.52	76.63
09	Male	.50	.31	.48	14.76	107.50
	Female	.34	.22	.31	1.97#	7.08
10	Male	.38	.27	.34	60.89	169.36
	Female	.34	.27	.27	21.93	21.46
11	Male	.41	.34	.34	10.17	9.37
	Female	-	-	-	-	-
12	Male	.38	.23	.29	20.57	32.18
	Female	-	-	-	-	-
13	Male	.29	.21	.23	8.69	11.73
	Female	-	-	-	-	-
14	Male	.52	.37	.48	5.40*	17.98
	Female	-	-	-	-	-
15	Male	.42	.23	.40	39.99	323.01
	Female	.41	.29	.35	29.96	58.27
16	Male	.52	.34	.49	9.53	38.71
	Female	.25	.20	.15	1.79	.97#

Table 5 (continued)

Group Nr.	Sample	R	Validity for: Educ Index	Validity for: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
17	Male	.41	.23	.38	96.99	610.22
	Female	.22	.21	.10	48.99	6.78
18	Male	.53	.31	.49	73.83	362.77
	Female	.32	.32	.00	38.30	.42#
19	Male	.32	.14	.32	.01#	23.21
	Female	.63	.60	.21	18.58	1.72#
20	Male	.37	.21	.34	21.97	107.80
	Female	-	-	-	-	-
21	Male	.33	.22	.27	41.59	73.08
	Female	-	-	-	-	-
22	Male	.53	.45	.46	18.57	23.39
	Female	-	-	-	-	-
23	Male	.55	.26	.53	9.54	96.12
	Female	-	-	-	-	-
24	Male	.22	.11	.20	1.57#	7.03
	Female	.45	.20	.42	3.17#	16.94
25	Male	.37	.23	.33	22.40	66.11
	Female	.27	.27	.01	13.28	.00#
26	Male	.47	.17	.46	11.95	34.87
	Female	.30	.29	.11	1.94#	.16#
27	Male	.40	.15	.40	1.42#	75.99
	Female	.07	.07	.03	.45#	.04#
28	Male	.52	.37	.45	23.75	46.07
	Female	-	-	-	-	-
29	Male	.27	.20	.24	14.02	25.64
	Female	-	-	-	-	-

Table 5 (continued)

Group Nr.	Sample	R	Validity for: Educ Index	Validity for: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
30	Male	.49	.27	.47	3.11#	40.85
	Female	.44	.44	-.09	5.81*	.09#
31	Male	.29	.27	.18	79.18	17.88
	Female	.34	.30	.21	15.58	5.74*
32	Male	.07	.06	-.02	1.16#	.45#
	Female	.23	.22	.08	5.33*	.29#
33	Male	.27	.15	.26	3.28#	34.59
	Female	-	-	-	-	-
34	Male	.28	.26	.12	87.34	11.04
	Female	.31	.29	.16	44.39	9.07
35	Male	.19	.14	.16	5.90*	9.06
	Female	.41	.22	.39	6.12*	45.96
36	Male	.30	.30	.03	35.99	.02#
	Female	.12	.12	.04	2.49#	.03#
37	Male	.30	.27	.19	102.70	30.56
	Female	.25	.18	.19	22.22	26.57
38	Male	.40	.36	.22	58.60	15.01
	Female	.34	.25	.27	9.05	11.64
39	Male	.28	.24	.21	132.42	71.33
	Female	-	-	-	-	-
40	Male	.41	.28	.38	20.73	91.95
	Female	.42	.29	.38	20.56	52.22
41	Male	.40	.31	.36	36.36	75.57
	Female	.43	.37	.33	43.39	27.55
42	Male	.47	.35	.42	74.24	163.82
	Female	.37	.30	.29	36.36	30.99

Table 5 (continued)

Group Nr.	Sample	R	Validity for: Educ Index	Validity for: Selector AI	F for Contribution of: Educ Index	F for Contribution of: Apt. Index
43	Male	.45	.36	.36	21.93	22.97
	Female	.44	.05	.44	.40#	10.86
44	Male	.18	.17	.10	7.35	.47#
	Female	.21	.21	.08	2.46#	.07#
45	Male	.49	.41	.41	21.30	20.72
	Female	.53	.38	.48	7.56	19.67

\# Not significant.
* Significant at the .05 level. All others are significant at the .01 level.

Table 6. Multiple R's and Tests of Hypotheses

Group Nr.	N	Class	Regression[1] R_1	R_2	R_3	R_4	F for Hypothesis[2] H_1	H_2	H_3
01	290	Race	.532	.527	.527	.473	3.85**	.47	.86
		Sex	.481	.481	.481	.473	.92	-	-
03	214	Race	.514	.511	.513	.456	2.64*	.10	.49
		Sex	.462	.462	.460	.456	.49	-	-
04	374	Race	.476	.476	.475	.358	7.84**	.16	.43
		Sex	.375	.361	.360	.358	1.84	-	-
05	905	Race	.393	.414	.416	.356	4.78**	.64	.00
		Sex	.358	.358	.358	.356	.69	-	-
06	567	Race	.402	.401	.398	.383	15.80**	.45	.03
		Sex	.386	.384	.384	.383	.37	-	-
08	2083	Race	.493	.492	.490	.478	6.58**	1.99	1.24
		Sex	.478	.478	.478	.478	.45	-	-
09	612	Race	.509	.507	.495	.478	4.16**	2.83*	.74
		Sex	.484	.481	.482	.478	1.51	-	-
10	2464	Race	.389	.388	.387	.365	8.54**	.87	.72
		Sex	.372	.371	.371	.365	4.66**	.86	1.41
11	178	Race	.443	.439	.429	.384	1.72*	-	-
		Sex	.433	.432	.432	.384	2.84	.01	.09
12	303	Race	.416	.404	.389	.379	1.77	-	-
		Sex	-	-	-	-	-	-	-
13	302	Race	.372	.359	.351	.301	2.76*	1.33	1.74
		Sex	.357	.349	.342	.301	4.20**	1.80	1.93
14	103	Race	.595	.585	.573	.548	1.34	-	-
		Sex	.568	.568	.553	.548	1.06	-	-
15	2737	Race	.430	.430	.427	.421	4.47**	2.01*	.17
		Sex	.428	.427	.425	.421	6.35**	3.51*	.33
16	221	Race	.543	.527	.524	.501	2.24*	1.59	2.61
		Sex	.522	.505	.504	.501	2.12	-	-
17	5736	Race	.420	.420	.417	.368	47.65**	4.70**	.70
		Sex	.394	.370	.375	.368	44.31**	49.52**	123.49**

Table 6 (Continued)

Group Nr.	N	Class	Regression[1] R_1	R_2	R_3	R_4	F for Hypothesis[2] H_1	H_2	H_3
18	1763	Race	.515	.513	.512	.481	13.14**	1.55	1.44
		Sex	.505	.491	.490	.481	18.10**	17.83**	32.85**
19	295	Race	.418	.417	.410	.351	2.98**	.61	.14
		Sex	.397	.392	.354	.351	3.97**	5.62**	1.43
20	1008	Race	.429	.427	.418	.399	5.03**	2.83*	1.10
		Sex	-	-	-	-	-	-	-
21	1084	Race	.376	.374	.368	.329	6.98**	1.85	1.07
		Sex	-	-	-	-	-	-	-
22	209	Race	.586	.577	.572	.531	3.14**	1.25	1.50
		Sex	.560	.560	.560	.531	3.17*	-	-
23	282	Race	.577	.575	.569	.553	1.90	-	-
		Sex	-	-	-	-	-	-	-
24	269	Race	.320	.294	.294	.283	1.08	-	-
		Sex	.315	.304	.297	.283	1.85	-	-
25	827	Race	.411	.408	.391	.317	11.18**	4.01**	1.09
		Sex	.354	.340	.336	.317	7.72**	5.89**	9.43**
26	1151	Race	.442	.442	.437	.426	3.30**	1.78	.14
		Sex	.452	.442	.439	.426	10.76**	8.08**	16.17**
27	590	Race	.400	.393	.396	.346	3.61*	4.14*	9.66**
		Sex	.368	.348	.351	.346	14.73**	.09	1.85
28	251	Race	.565	.565	.563	.521	2.87*	.19	.02
		Sex	-	-	-	-	-	-	-
29	711	Race	.420	.414	.419	.270	14.73**	.09	1.85
		Sex	.278	.278	.278	.270	1.15	-	-
30	223	Race	.495	.489	.465	.439	2.50*	2.08	.82
		Sex	.484	.460	.444	.439	3.99**	5.32**	6.67*
31	1546	Race	.409	.408	.400	.306	22.70**	3.61**	.83
		Sex	.308	.307	.307	.306	.62	-	-
32	401	Race	.267	.255	.227	.106	4.23**	2.07	1.33
		Sex	.269	.269	.262	.106	2.95*	.79	.04

Table 6 (Continued)

Group Nr.	N	Class	Regression[1] R_1	R_2	R_3	R_4	F for Hypothesis[2] H1	H_2	H_3
33	646	Race	.417	.415	.412	.269	13.07**	.83	.54
		Sex	.269	.269	.269	.269	.02	-	-
34	1869	Race	.334	.334	.330	.286	10.27**	1.36	.21
		Sex	.289	.288	.287	.286	.88	-	-
35	858	Race	.332	.318	.307	.278	5.33**	3.93**	4.52**
		Sex	.310	.287	.280	.278	5.98**	8.28**	13.26**
36	551	Race	.323	.323	.319	.254	4.04**	.40	.09
		Sex	.265	.265	.255	.254	1.16	-	-
37	2637	Race	.338	.338	.333	.279	18.17**	2.75*	.45
		Sex	.294	.294	.292	.279	8.26**	1.59	.87
38	648	Race	.472	.471	.464	.378	11.00**	1.63	.37
		Sex	.389	.389	.386	.378	2.20	-	-
39	3438	Race	.395	.394	.391	.279	53.27**	3.66**	2.65
		Sex	-	-	-	-	-	-	-
40	1348	Race	.475	.474	.473	.423	13.40**	.65	.43
		Sex	.441	.439	.440	.423	8.73**	.92	2.51
41	1366	Race	.525	.525	.519	.405	34.83**	2.63*	.38
		Sex	.417	.416	.415	.405	5.61**	1.98	2.31
42	1903	Race	.525	.524	.522	.434	37.95**	1.70	.92
		Sex	.439	.436	.436	.434	3.14*	2.35	6.12*
43	300	Race	.449	.443	.442	.432	.94	-	-
		Sex	.455	.455	.439	.432	2.56	-	-
44	402	Race	.290	.284	.268	.203	3.07**	1.29	.70
		Sex	.247	.247	.247	.203	2.84*	.06	.09
45	320	Race	.538	.531	.520	.496	3.18**	2.04	1.65
		Sex	.507	.506	.505	.496	1.54	-	-

[1] Multiple correlations for prediction of Final School Grade. Predictors are:

R_1 - Education Index (EI), Selector AI (SI), Classification Membership (CI), and interactions (or products) of EI with CI and of SI with CI.

R_2 - EI, SI, and CI

R_3 - EI, SI, Cl x EI, Cl x SI

R_4 - EI and SI

[2] Hypotheses are as follow:

H_1 = Classification makes no contribution to prediction (cf R_1^2 and R_4^2

H_2 = Classification regression slopes are homogeneous (cf R_1^2 and R_3^2)

H_3 = Classification regression intercepts are homogeneous (cf R_1^2 and R_2^2)

** Significant at the .01 level.
* Significant at the .05 level.

References

Campbell, J. Testing of Culturally different groups. Research Bulletin 64-34. Princeton, N.J.: Educational Testing Service, 1964.

Gordon, M.A. A study in the applicability of the same minimum qualifying scores for technical schools to White males, WAF, and Negro males. Technical Report 53-34, Lackland AFB, Tex.: Human Resources Research Center, November 1953.

Guinn, N., Tupes, E.C., & Alley, W.E. Demographic differences in aptitude test performance. AFHRL-TR-70-15. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, May, 1970(a)

Guinn, N., Tupes, E.C., & Alley, W.E. Cultural supgroup differences in the relationships between Air Force aptitude composites and training criteria. AFHRL-TR-70-35. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, September 1970(b).

Kirkpatrick, J.J., Even, R.B., Barrett, R.S., & Katzell, R.A. Testing and fair employment: Fairness and validity of personnel tests for different ethnic groups. New York: New York University Press; London: University of London Press Limited, 1968.

Lopez, F.M., Jr. Current problems in test performance of job applicants: I. Personnel Psychology, 1966, 19, 10-18.

Tenopyr, M.L. Race and socio-economic status as moderators in predicting machine-shop training success. Paper presented at the annual convention of the American Psychological Association, Washington, D.C., 1967.

Prediction Of Enlisted Tenure - (Two Years)
The POET-2 Model

by
William A. Sands

Navy Personnel Research and Development Center[1]
San Diego, California 92152

INTRODUCTION

Slide #1

The research discussed in this paper was initiated by a request from the Navy Recruiting Command and was accomplished under the Bureau of Naval Personnel Management Support Program, Testing and Classification, Recruit Prediction.

Slide #2

Today I will cover the background of the study, outline the problem and purposes of the work, and describe the development and evaluation of the POET-2, or Prediction of Enlisted Tenure, Model. The cohort of sailors will be described and the results of statistical and management-oriented analyses will be presented.

Background

Slide #3

In 1960, a research program was initiated at the Navy Medical Neuropsychiatric Research Unit in San Diego, CA. A total of 11,008 men were followed throughout their first enlistment to determine the effectiveness of their military performance. Effective service was operationally defined as satisfactory completion of obligated service and a recommendation for reenlistment by the man's Commanding Officer. A major end product of this longitudinal research program was the original Odds for Effectiveness Table. This OFE-1 table was designed to estimate the probability that a male applicant with no prior service would render satisfactory naval service as a function of preservice attributes. The OFE-1 table estimated the probability of success based on the following background information: (1) mental group, determined by an aptitude test score; (2) number of years of school completed; (3) number of expulsions and/or suspensions from school; and (4) number of arrests, excluding traffic violations.

---

[1]The views expressed herein are those of the author and do not necessarily reflect those of the Navy Department.

This last variable influenced predictions only for certain mental groups. The OFE-1 table was updated in 1968 and operationally implemented for screening recruit applicants during the first part of CY 1973. Somewhat later, a problem in using the OFE-1 table was encountered by recruiting personnel in the field. For various reasons, arrest records became increasingly difficult for the recruiter to obtain. Some law enforcement agencies refused to supply such information. This created a problem for the recruiter who had been instructed to determine an OFE-1 score for each nonprior service enlisted male applicant. Various procedures for handling this missing data problem were considered, for example, assuming zero arrests in the absence of contrary evidence. However, these "convenient" solutions were inadequate from an administrative and/or technical standpoint. As a result, the Navy Personnel Research and Development Center was requested to formulate a revised OFE table which would not require arrest information.

Slide #4

Due to the operational demand for an immediate solution, a new cohort of enlistees could not be followed throughout their first enlistment. Instead the data originally collected in 1960-61 were reevaluated and a new OFE table was constructed. This revised OFE table was submitted to the Navy Recruiting Command with the recommendation that it supersede the original OFE table. As of 1 October 1975, the revised table (OFE-2) was incorporated into the Navy Recruiting Manual, Enlisted and became operational.

Although the OFE-2 table represented an interim solution to a pressing operational problem, the data used for constructing the table are about fifteen years old. In addition, the data were collected during a time when enlistment in the Navy was influenced to some extent by the pressure of the draft.

Purpose

The global purpose of this research is to develop effective tools to aid Navy recruiters in screening applicants for enlistment. These tools should reduce the large financial, social, and personal costs of erroneous acceptances. The specific objective is to develop an instrument which can provide the recruiter with an estimate of an applicant's probability of surviving for two years. Those applicants with low probability could be screened out, resulting in a decrease in premature attrition.

Slide #5

The POET-2 model is designed to provide these estimates as a function of: (1) years education completed; (2) mental group; (3) age at active duty base date; and (4) number of primary dependents.

PROCEDURE

Sample

Slide #6

Data were available on essentially all nonprior service enlisted male accessions with an active duty base date in CY 1973. For the sample of 68,616 men descriptive information was available on a variety of characteristics. The total sample was divided into two subsamples: a majority group composed entirely of Caucasians; and a minority group composed of Negroes, American Indians, Malayans, and Mongolians. The Caucasians make up 89% of the total group and the Negroes make up 91% of the minority sample.

Criterion

Criterion data on the cohort were obtained as of June 1975. Persons with an active duty base date in January 1973 would have served thirty months while individuals with an active duty base date in December 1973 would have served eighteen months, producing a criterion based upon a two year median length of service.

The specific outcomes for the sample of men were grouped into three categories: (1) survivors, (2) losses and (3) indeterminates. The indeterminates, for example, death, were removed from the sample prior to the statistical analyses, leaving a dichotomous criterion of survival (72%) versus loss (28%) based upon a total of 68,616 men.

Slide #7

This slide provides some information on the losses by type of discharge. The most common type of discharge was honorable, accounting for about 49% of the total losses. Desertion was an example of the type of loss included under the "other" category.

Predictors

As previously indicated, four types of information were chosen to use in predicting the probability of survival: (1) years of education completed (GED was considered as completion of twelve years); (2) mental group as defined by the Armed Forces Qualification Test (AFQT) score level, based upon the Short Basic Test Battery (SBTB); (3) age at active duty base date; and (4) number of primary dependents. These four variables were chosen because of their demonstrated predictive value and because the information is collected routinely by Navy recruiters for all enlisted male applicants.

Slide #8

The frequency distribution for each group on years of education completed is shown in this slide. Approximately 71% of the total cohort were high school graduates or had earned a GED.

Slide #9

The mental group frequency distribution for each group is shown in this slide.

The Navy Recruiting Command uses a two-by-two table which combines education and mental group for the purpose of formulating and evaluating recruiting policy. The four cells of the table are defined as follows:

A. School eligible and high school graduate

B. School eligible and nonhigh school graduate

C. Nonschool eligible and high school graduate

D. Nonschool eligible and nonhigh school graduate

School eligibility is operationally defined in terms of mental group. Persons in mental groups I, II, and the top portion of III are classified as school eligible while persons in lower group III and group IV are considered nonschool eligible. While the percentage of high school graduates is close for the majority and minority groups, the percentage of persons qualifying as school eligible is markedly different. Specifically, while 71% of the majority sample was qualified as school eligible, only 35% of the minority group was considered school eligible.

Slide #10

The frequency distribution for age at active duty base date is shown in this slide. The proportion of persons entering at age seventeen is greater for the majority group than for the minority group.

Slide #11

This slide shows the frequency distribution for the number of primary dependents. The proportion of each group with one or more primary dependents is very similar.

Analyses

Whenever a multiple regression analysis is performed, two conflicting objectives exist. The first, and most important, objective is to obtain an equation with weights which are as stable as possible. This is achieved by using all available data in the development of the regression weights, the procedure followed in constructing the POET-2 model. The other objective is to evaluate the accuracy or effectiveness of the equation. Since an unbiased estimate of the accuracy can be obtained only on a sample of cases which was not used in the development of the regression weights, a holdout sample must be removed from the total sample prior to developing the weights, thereby decreasing the stability of the weights, the first objective.

To obtain cross-validation evidence for the POET-2 model, the total sample was equally divided into a development and an evaluation sample. Using stratified random selection, this division was done in such a way as to insure comparable survival rates. A regression equation was constructed in the development group and applied to the evaluation group. Another equation was formed on the evaluation group and applied to the development group. This procedure is known as double cross-validation.

Slide #12

The effectiveness of a personnel selection instrument is a function of three considerations: (1) the base rate; (2) the validity; and (3) the selection ratio. These concepts can be illustrated using a scatter diagram as shown in this slide. Each point in the elipse represents an observation or persón. The location of the point is determined by the intersection of his predictor and criterion scores.

Slide #13

If the criterion is a dichotomous variable such as survival versus loss and if a predictor cutting score is established, four decision-outcome combinations result as shown in this slide. Erroneous rejections represent persons who were rejected but who would have been survivors if given the opportunity. Correct rejections represent persons rejected who would have attrited if they had been selected. Correct acceptances are persons selected who survive. Finally, erroneous acceptances are persons selected who subsequently attrite.

The base rate represents the proportion of survivors in the total group; i.e., persons above the horizontal dashed line. The validity of a selection instrument is the correlation between scores on that instrument and the criterion. This correlation expresses the degree of linear association between the predictor and criterion. If there were no relationship, the figure in this slide would be a circle. If the relationship were perfect, the elipse would collapse and leave a straight line allowing perfect forecasts of the criterion from the predictor. The selection ratio is the proportion of persons selected at a particular cutting score on the predictor as represented by the dashed vertical line, i.e., the number of correct acceptances plus the number of erroneous acceptances divided by the total number of persons.

The factor over which the personnel manager has the most control is the selection ratio. If it is desired to reduce attrition, he lowers the selection ratio, or proportion selected, by raising the cutting score. Although this action will reduce erroneous acceptances, it will have two negative consequences: (1) the number of correct acceptances will be reduced; and (2) the number of erroneous rejections will be increased. If the manager wants to avoid denying acceptance to persons who would survive, he can raise the selection ratio by lowering the cutting score. This action will reduce the number of erroneous rejections, but only at the price of increasing the attrition rate.

In summary, the cutting score can be shifted to accomplish any desired policy. However, reducing one type of erroneous selection decision can only be accomplished by increasing the other type of erroneous decision. The most appropriate strategy would vary with the problem and would be a matter of policy.

RESULTS

Slide #14

Survival Rates by Predictor

This slide presents the survival rates for the majority, minority, and total samples for selected variables. Education has been divided into four categories: (1) nine years or less; (2) ten or eleven years; (3) twelve years or GED; and (4) thirteen years or more. The survival rate increases as years of education increase.

Mental group is presented as five distinct categories. Examination of the survival rates for the mental groups shows the expected progression except for the bottom two groups. Perhaps the members of mental group IV who were allowed to enlist were superior to those in lower mental group III in other respects.

Age at active duty base date was dichotomized. As shown, persons enlisting at age seventeen have a lower rate of survival than do persons eighteen or older.

The number of primary dependents was dichotomized. Persons with no primary dependents evidenced a higher survival rate than did persons with one or more primary dependents.

As shown, the survival rate differs markedly for the racial groups. Rank-ordering the groups from highest to lowest survival rate shows Malayans on top, followed by Mongolians, Caucasians, Negroes and finally, American Indians as being the highest risk group.

Slide #15

Model Specification

The POET-2 model is based upon multiple regression analysis. The prediction equation for a multiple linear regression analysis model is shown in this slide.

Slide #16

The POET-2 model is shown in this slide. Each of the predictors is a dummy variable which takes on a value of one if it is a characteristic of the applicant and zero otherwise. One interval of each predictor acts to define a control group for the equation. These control group intervals are indicated by the asterisks. The weight that they receive is reflected in the constant. The control

group for the POET-2 model equation has twelve years education, scores in mental group II, is eighteen or older and has no primary dependents. Examination of the equation reveals that it includes nine predictors and a constant.

Slide #17

The application of the POET-2 model can be illustrated using a hypothetical applicant with the following characteristics: twelve years education; mental group II; age eighteen; and one primary dependent. This information is used to establish the predictor column of ones and zeroes. Multiplying this column by the regression weight column and summing the results produces the best estimate of the chances of this hypothetical applicant surviving for two years—77 chances in 100.

Chances of Survival

Slide #18

To eliminate the necessity of substituting the appropriate values into the regression equation and computing an estimated probability of survival for each male nonprior service enlisted applicant, all configurations were computed and tabled. There were 80 unique combinations of predictors for the POET-2 model. These predictions were organized into a single table designed for application to all male nonprior service prospective enlistees. Scores produced by this range from 93 to 29, thus permitting considerable flexibility in setting cutting scores to accommodate fluctuation in manpower supply and demand.

Validation and Cross-Validation

Slide #19

The multiple correlation coefficient for the POET-2 model computed on the total sample of 68,616 men was .31. As previously indicated, cross-validation is necessary to ensure the stability of the predictor weights. The total sample was divided equally into two groups: A development sample and an evaluation sample. A multiple regression equation was constructed on the development sample ($R = .31$) and applied to the evaluation sample, giving a cross-validity of $r = .30$. A second equation was constructed on the evaluation sample ($R = .30$) and applied to the development sample, producing a cross-validity of $r = .31$. The results of this double cross-validation indicated that the POET-2 model produces stable predictions.

Management-Oriented Information

Slide #20

The statistical results which have been presented are necessary but not sufficient. Validity coefficients and related indices do not provide any direct information on the consequences of applying

a selection procedure. This table illustrates the consequences of applying the POET-2 model to the sample at alternative cutting scores.

The score in the lefthand column is the probability of surviving two years as estimated by the POET-2 model. These scores are the numbers contained in the chances for survival table designed for use by recruiters in the field.

The selection ratio is the proportion selected. The survival rate is the number of survivors divided by the number accepted at a particular cutting score. The hit rate represents the proportion of correct selection decisions, i.e., correct rejections and correct acceptances.

The next four columns show the number of each of the four decision-outcome combinations. The next two columns show the total number rejected and total number accepted. Finally, the last four columns show the number of persons accepted who are in each of the four recruit quality categories.

To illustrate the use of these tables, let us examine the consequences of applying a cut score of 54. This strategy would select 83.5% of the original sample. The survival rate for the group selected would be 76.7%. The percent correct decisions or hit rate would be 72.6%. Fifty-eight hundred twenty-six persons would be correctly rejected while 5,515 would have been erroneously rejected. Forty-three thousand nine hundred and fifty-five men would have been correctly accepted. At the same time 13,320 men who would be accepted would attrite. A total of 11,341 men would be rejected while 57,275 would be accepted. Of this group of selectees 36,581 would be in recruit quality category A, 8,500 in B, 12,194 in C, and none in category D.

Using this table, the manager of the personnel system can evaluate the consequences of employing alternative cutting scores in terms of various tradeoffs and can weight these consequences in the light of the current supply and demand picture for nonprior service enlisted males.

Conclusions

In conclusion, the POET-2 model represents a tool which can be used effectively by Navy recruiters for screening nonprior service enlisted male applicants. The statistical evidence shows that the model will produce reasonably accurate predictions. The availability of the analyses showing consequences of using alternative cutting scores enhances the flexibility of using the POET-2 model for the personnel manager.

Slide #1

PREDICTION OF ENLISTED TENURE TWO YEARS

THE POET-2 MODEL

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CA 92152

Slide #2

OUTLINE

- INTRODUCTION
  - BACKGROUND
  - PURPOSE
- PROCEDURE
  - SAMPLE
  - CRITERION
  - PREDICTORS
  - ANALYSES
- RESULTS
  - SURVIVAL RATES BY PREDICTOR
  - MODEL SPECIFICATION
  - CHANCES OF SURVIVAL
  - VALIDATION AND CROSS-VALIDATION
  - MANAGEMENT-ORIENTED INFORMATION
- CONCLUSIONS AND RECOMMENDATIONS

Slide #3

The Original Odds for Effectiveness Table

The OFE-1 Model

CRITERION:

1. Satisfactory Completion of Obligated Service
2. Recommendation for Reenlistment

PREDICTORS:

1. Mental Group
2. Years School Completed
3. Expulsions/Suspensions
4. Arrests

Slide #4

The Revised Odds for Effectiveness Table

The OFE-2 Model

CRITERION:

1. Satisfactory Completion of Obligated Service
2. Recommendation for Reenlistment

PREDICTORS:

1. Mental Group
2. Years School Completed
3. Expulsions/Suspensions

Slide #5

The Prediction of Enlisted Tenure-Two Years

The POET-2 Model

CRITERION:

Survival for First Two Years

PREDICTORS:

1. Years School Completed
2. Mental Group
3. Age at ADBD
4. Primary Dependents

Slide #6

Frequency Distribution for Racial Group Membership

Racial Group	Majority (N=61,116) N	%	Minority (N=7,500) N	%	Total (N=68,616) N	%	
Caucasion	61116	100.0	---	---	61116	89.07	89.07
Negro	---	---	6795	90.60	6795	9.90	10.93
American Indian	---	---	270	3.60	270	0.39	
Maylayan	---	---	292	3.89	292	0.43	
Mongolian	---	---	143	1.91	143	0.21	

898

1063

1064

Slide #7

Frequency of Losses by Discharge Type

Outcome Type	Majority (N=16,587) N	%	Minority (N=2,559) N	%	Total (N=19,146) N	%
Honorable Discharge	8115	48.92	1197	46.78	9312	48.64
General Discharge	6723	40.53	1159	45.29	7882	41.17
Undesirable Discharge	950	5.73	76	2.97	1026	5.36
Bad Conduct Discharge	32	0.19	4	0.16	36	0.19
Other	767	4.62	123	4.81	890	4.65

696

1065

1066

Slide #8

Frequency Distribution for Years Education Completed

Years Education Completed	Majority (N=61,116) N	%		Minority (N=7,500) N	%		Total (N=68,616) N	%	
6-	2	0.00		1	0.01		3	0.00	
7	27	0.04		2	0.03		29	0.04	
8	259	0.42	3.34	11	0.15	1.55	270	0.39	3.14
9	1753	2.87		102	1.36		1855	2.70	
10	7608	12.45		698	9.31		8306	12.11	
11	7824	12.80	25.25	1554	20.72	30.03	9378	13.67	25.77
12 or GED	38833	63.54	63.54	4543	60.57	60.57	43376	63.22	63.22
13	2490	4.07		265	3.53		2755	4.02	
14	1513	2.48		195	2.60		1708	2.49	
15	352	0.58	7.87	69	0.92	7.85	421	0.61	7.87
16	429	0.70		59	0.79		488	0.71	
17+	26	0.04		1	0.01		27	0.04	

970

Slide #9

Frequency Distribution for Mental Groups

Mental Group	Majority (N=61,116) N	%	Minority (N=7,500) N	%	Total (N=68,616) N	%
I	1943	3.18	20	0.27	1963	2.86
II	23541	38.52	698	9.31	24239	35.33
III-U	18104	29.62	1896	25.28	20000	29.15
III-L	16069	26.29	4288	57.17	20357	29.67
IV	1459	2.39	598	7.97	2057	3.00

971

1069

1070

Slide #10

Frequency Distribution for Age at Active Duty Base Date

Age	Majority (N=61,116) N	%		Minority (N=7,500) N	%		Total (N=68,616) N	%	
17	17551	28.72	28.72	1703	22.71	22.71	19254	28.06	28.06
18	22014	36.02		2486	33.15		24500	35.71	
19	11692	19.13		1522	20.29		13214	19.26	
20	5083	8.32		783	10.44		5866	8.55	
21	2260	3.70		401	5.35		2661	3.88	
22	1253	2.05	71.28	281	3.75	77.29	1534	2.24	71.94
23	606	0.99		151	2.01		757	1.10	
24	278	0.45		71	0.95		349	0.51	
25	145	0.24		39	0.52		184	0.27	
≥26	234	0.38		63	0.84		297	0.43	

972

1071

1072

Slide #11

Frequency Distribution for Number of Primary Dependents

Number Primary Dependents	Majority (N=61,116) N	%		Minority (N=7,500) N	%		Total (N=68,616) N	%	
0	57387	93.90	93.90	7019	93.59	93.59	64406	93.86	93.86
1	2193	3.59		238	3.17		2431	3.54	
2	1414	2.31	6.10	231	3.08	6.41	1645	2.40	6.14
≥3	122	0.20		12	0.16		134	0.20	

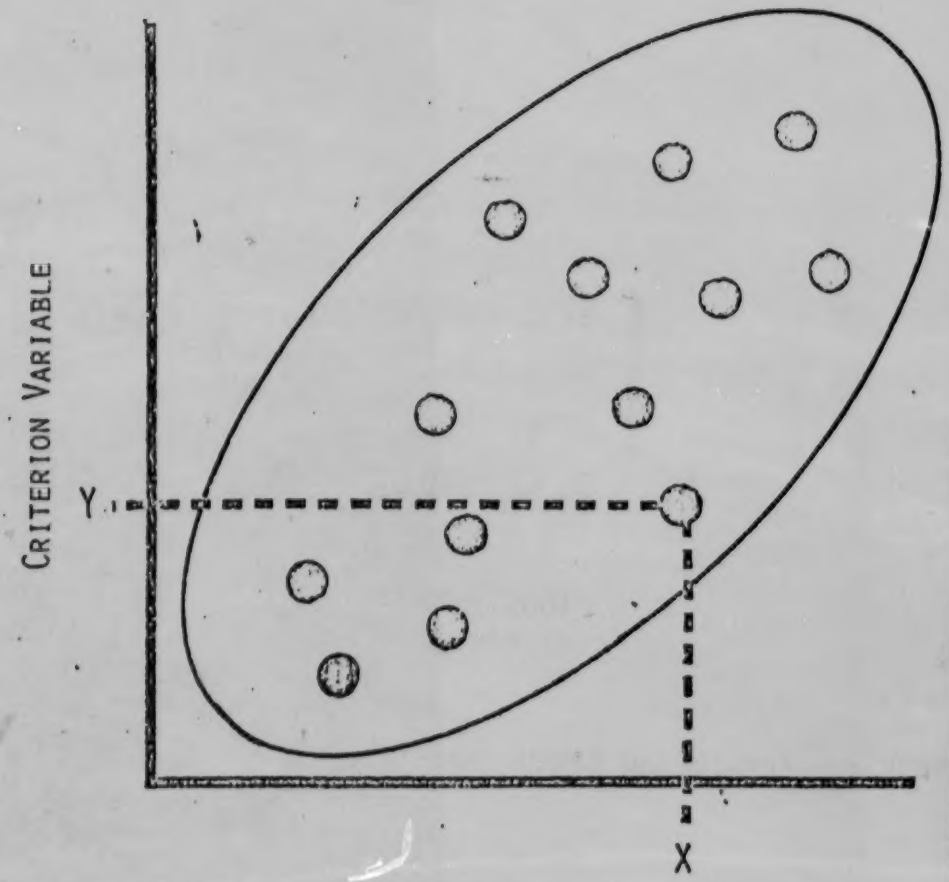
973

1073

1074

Slide #12

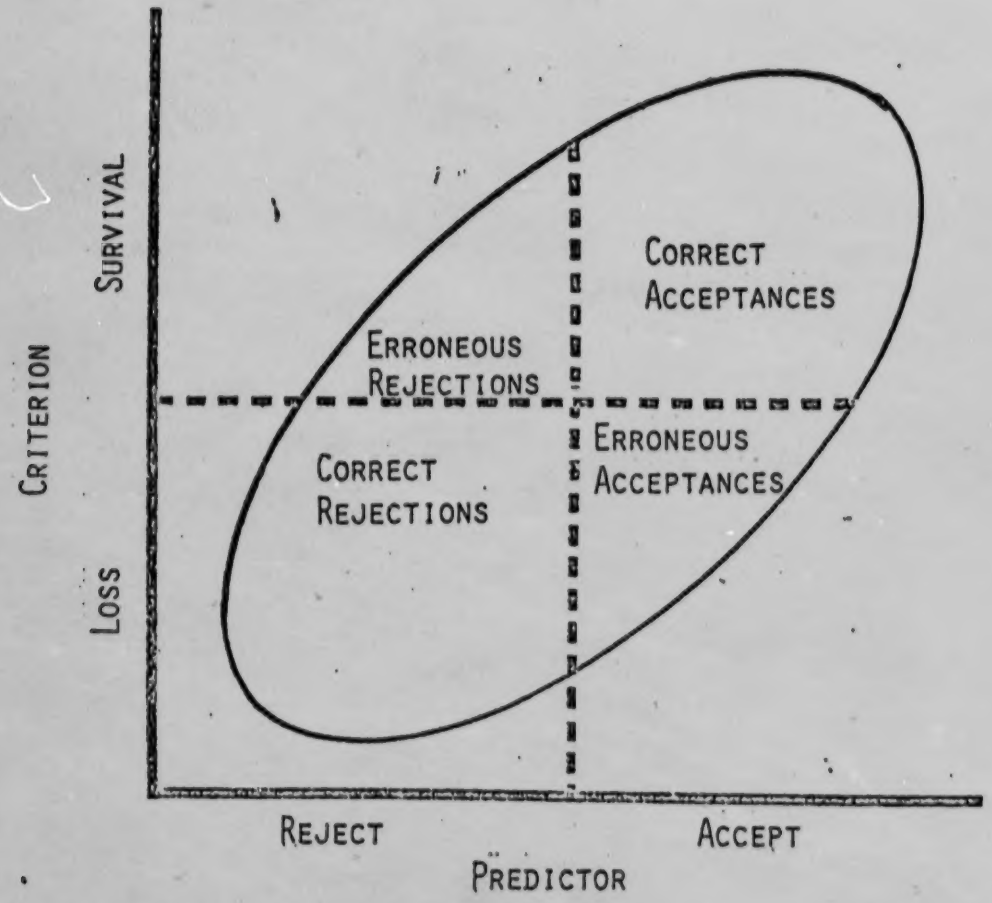
BIVARIATE SCATTER DIAGRAM



974

Slide #13

FOUR SELECTION DECISION-OUTCOME COMBINATIONS



975

1076

1077

Slide #14

Survival Rates by Predictor Variable

Predictor Variable	Majority (N=61,116)	Minority (N=7,500)	Total (N=68,616)	
Years Education Completed				
≤ 9	.429	.448	.430	
10-11	.545	.480	.537	
12 or GED	.799	.731	.792	
≥13	.874	.829	.869	
Mental Group				
I	.878	.950	.878	
II	.813	.795	.812	
III-U	.713	.685	.710	
III-L	.611	.618	.613	
IV	.662	.696	.672	
Age at ADBD				
17	.609	.564	.605	
≥18	.777	.687	.766	
Primary Dependents				
0	.731	.661	.724	
≥1	.686	.620	.679	
Racial Group				
Caucasian	.729	---	.729	.729
Negro	---	.644	.644	
American Indian	---	.641	.641	
Malayan	---	.901	.901	.659
Mongolian	---	.888	.888	

Slide #15

General Multiple Linear Regression Model

$$\hat{Y} = W_0 U + \sum_{J=1}^{K} W_J X_J$$

Where: $\hat{Y}$ = The criterion prediction

W_0 = The intercept or constant

U = The Unit vector

W_J = The least squares raw score regression weight for the J^{TH} predictor ($J = 1, 2, ---, K$).

X_J = The value for the J^{TH} predictor ($J = 1, 2, ---, K$)

Slide #16

POET-2 Model

Dummy Variables	Regression Weight
Constant	.84544
ED9^{-}	-.29614
ED10-11	-.21069
*ED12	-----
ED13^{+}	.04869
MGI	.03291
*MGII	-----
MGIII-U	-.04643
MGII-L	-.10452
MGIV	-.13914
AGE17	-.04475
*AGE18^{+}	-----
*NPDEP	-----
PDEP	-.07114

*Dummy predictors subsumed under the constant.

Slide #17

Application of the POET-2 Model to a Hypothetical Applicant

Description:

Twelve Years Education
Mental Group II
Age 18
One Primary Dependent

Dummy Variables	Predictor Value	Regression Weight	(Value)(Weight)
Constant	1	.84544	.84544
ED9$^-$	0	-.29614	.00000
ED10-11	0	-.21069	.00000
ED13$^+$	0	.04869	.00000
MGI	0	.03291	.00000
MGII-U	0	-.04643	.00000
MGIII-L	0	-.10452	.00000
MGIV	0	-.13914	.00000
AGE17	0	-.04475	.00000
PDEP	1	-.07114	-.07114
			.77430

Probability of Survival = .77

Slide #18

PREDICTION OF ENLISTED TENURE - TWO YEARS
THE POET-2 MODEL

Chances of Survival

Mental Group	Primary Dependents	9- Years Education		10-11 Years Education		12 Years Education*		13+ Years Education	
		Age 17	Age 18+	Age 17	Age 18+	Age 17	Age 18+	Age 17	Age 18+
I	No	54	58	62	67	83	88	88	93
	Yes	47	51	55	60	76	81	81	86
II	No	50	55	59	63	80	85	85	89
	Yes	43	48	52	56	73	77	78	82
III-U	No	46	50	54	59	75	80	80	85
	Yes	39	43	47	52	68	73	73	78
III-L	No	40	44	49	53	70	74	74	79
	Yes	33	37	41	46	63	67	67	72
IV	No	37	41	45	50	66	71	71	75
	Yes	29	34	38	42	59	64	64	68

*GED is considered as 12 years education.

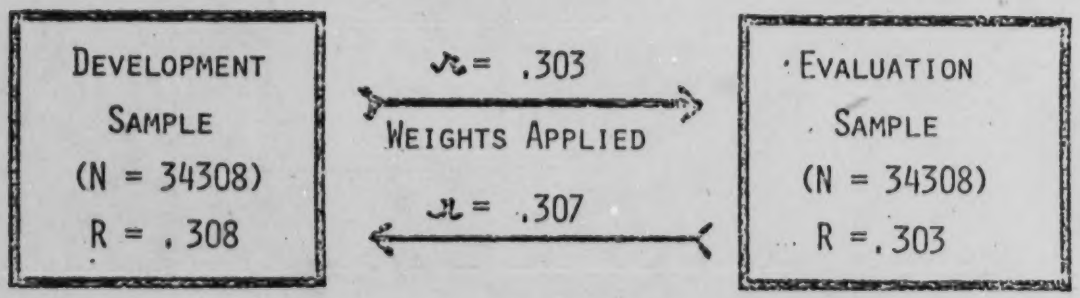
086

1082

1083

Slide #19

Double Cross-Validation Results for the POET-2 Model



Slide #20-A

- CONSEQUENCES OF USING THE PDET-2 MODEL FOR TOTAL GROUP -

CUT SCORE	SELECT. RATIO	SURVIV. RATE	HIT RATE	NUMBER CORRECT REJECT.	NUMBER ERROR. REJECT.	NUMBER CORRECT ACCEPT.	NUMBER ERROR. ACCEPT.	TOTAL NUMBER REJECT.	TOTAL NUMBER ACCEPT.	NUMBER ACCEPTED BY RECRUIT QUALITY CATEGORY: HSG SE (A)	NHSG SE (B)	HSG NSE (C)	NHSG NSE (D)
99	0.0	1.000	0.279	19146	49470	0	0	68616	0	0	0	0	0
98	0.0	1.000	0.279	19146	49470	0	0	68616	0	0	0	0	0
97	0.0	1.000	0.279	19146	49470	0	0	68616	0	0	0	0	0
96	0.0	1.000	0.279	19146	49470	0	0	68616	0	0	0	0	0
95	0.0	1.000	0.279	19146	49470	0	0	68616	0	0	0	0	0
94	0.0	1.000	0.279	19146	49470	0	0	68616	0	0	0	0	0
93	0.008	0.904	0.286	19093	48969	501	53	68062	554	554	0	0	0
92	0.008	0.904	0.286	19093	48969	501	53	68062	554	554	0	0	0
91	0.008	0.904	0.286	19093	48969	501	53	68062	554	554	0	0	0
90	0.008	0.904	0.286	19093	48969	501	53	68062	554	554	0	0	0
89	0.049	0.884	0.317	18757	46499	2971	389	65256	3360	3360	0	0	0
88	0.060	0.884	0.320	18638	45582	3888	508	64220	4396	4396	0	0	0
87	0.064	0.884	0.320	18638	45582	3888	508	64220	4396	4396	0	0	0
86	0.065	0.884	0.329	18625	45507	3963	521	64132	4484	4484	0	0	0
85	0.282	0.853	0.479	16306	32938	16532	2840	49244	19372	19372	0	0	0
84	0.282	0.853	0.479	16306	32938	16532	2840	49244	19372	19372	0	0	0
83	0.284	0.854	0.480	16290	32818	16652	2856	49108	19508	19508	0	0	0
82	0.290	0.854	0.484	16245	32490	16980	2901	48735	19881	19881	0	0	0
81	0.291	0.854	0.485	16227	32414	17056	2919	48641	19975	19975	0	0	0
80	0.471	0.835	0.595	13810	22473	26997	5336	36283	32333	32333	0	0	0
79	0.478	0.835	0.599	13720	22071	27399	5426	35791	32825	32333	0	492	0
78	0.480	0.835	0.600	13701	21987	27483	5445	35688	32928	32436	0	492	0
77	0.496	0.833	0.610	13484	21126	28344	5662	34610	34006	33514	0	492	0
76	0.496	0.833	0.610	13482	21120	28350	5664	34602	34014	33522	0	492	0
75	0.531	0.827	0.626	12859	19349	30121	6287	32208	36408	35837	0	571	0
74	0.641	0.814	0.681	10959	13689	35781	8188	24647	43969	35837	0	8132	0
73	0.651	0.812	0.685	10733	13213	36257	8413	23946	44670	36538	0	8132	0
72	0.652	0.812	0.685	10724	13173	36297	8422	23897	44719	36538	0	8181	0
71	0.674	0.808	0.694	10267	12125	37345	8879	22392	46224	36538	0	9686	0
70	0.700	0.803	0.703	9683	10906	38564	9463	20589	48027	36538	0	11489	0
69	0.700	0.803	0.703	9683	10906	38564	9463	20589	48027	36538	0	11489	0
68	0.701	0.803	0.703	9668	10872	38598	9478	20540	48076	36581	0	11495	0
67	0.707	0.801	0.705	9503	10574	38896	9643	20077	48539	36581	20	11938	0
66	0.709	0.801	0.706	9469	10470	39000	9681	19935	48681	36581	20	12080	0
65	0.709	0.801	0.706	9469	10470	39000	9681	19935	48681	36581	20	12080	0
64	0.711	0.801	0.706	9430	10429	39041	9716	19859	48757	36581	20	12156	0
63	0.727	0.798	0.712	9057	9698	39772	10089	18755	49861	36581	1088	12192	0
62	0.727	0.798	0.712	9048	9687	39783	10098	18735	49881	36581	1108	12192	0
61	0.727	0.798	0.712	9044	9687	39783	10098	18735	49881	36581	1108	12192	0
60	0.727	0.798	0.712	9047	9686	39784	10099	18733	49883	36581	1110	12192	0
59	0.782	0.752	0.721	7469	7483	41987	11677	14952	53664	36581	4889	12194	0
58	0.782	0.782	0.721	7468	7482	41988	11678	14950	53666	36581	4891	12194	0
57	0.782	0.732	0.721	7468	7482	41988	11678	14950	53666	36581	4891	12194	0
56	0.784	0.782	0.721	7411	7405	42065	11735	14816	53800	36581	5025	12194	0
55	0.784	0.782	0.721	7401	7392	42078	11745	14793	53823	36581	5048	12194	0
54	0.835	0.767	0.726	5826	5515	43955	13320	11341	57275	36581	8500	12194	0
53	0.885	0.753	0.726	4189	3829	45641	14957	8018	60598	36581	8500	12194	3323
52	0.888	0.752	0.726	4029	3655	45815	15117	7684	60932	36581	8834	12194	3323
51	0.888	0.752	0.726	4028	3655	45815	15118	7683	60933	36581	8835	12194	3323
50	0.893	0.751	0.727	3880	3484	45992	15266	7368	61248	36581	9055	12194	3418

982

1084A 01

1084 B

Slide #20-B

- CONSEQUENCES OF USING THE POET-2 MODEL FOR TOTAL GROUP -

CUT SCORE	SELECT. RATIO	SURVIV. RATE	HIT RATE	NUMBER CORRECT REJECT.	NUMBER ERROR. REJECT.	NUMBER CORRECT ACCEPT.	NUMBER ERROR. ACCEPT.	TOTAL NUMBER REJECT.	TOTAL NUMBER ACCEPT.	NUMBER ACCEPTED BY RECRUIT QUALITY CATEGORY			
										HSG SE (A)	NHSG SE (B)	HSG NSE (C)	NHSG NSE (D)
49	0.963	0.732	0.726	1433	1082	48308	17713	2515	66101	36581	9055	12194	8271
48	0.963	0.732	0.726	1433	1081	48309	17713	2514	66102	36581	9056	12194	8271
47	0.964	0.732	0.726	1398	1039	48431	17748	2437	66179	36581	9133	12194	8271
46	0.976	0.728	0.725	964	692	48778	19102	1656	66960	36581	9876	12194	8609
45	0.977	0.728	0.725	924	637	48833	18922	1561	67055	36581	9576	12194	8704
44	0.982	0.727	0.724	733	493	48977	18913	1226	67390	36581	9576	12194	9039
43	0.983	0.727	0.724	712	481	48989	18434	1193	67423	36581	9609	12194	9039
42	0.983	0.727	0.724	705	477	48993	18441	1182	67434	36581	9609	12194	9050
41	0.984	0.726	0.724	650	439	49031	18496	1089	67527	36581	9609	12194	9143
40	0.998	0.722	0.722	84	42	49428	19062	126	68490	36581	9609	12194	10106
39	0.998	0.722	0.722	77	37	49433	19069	114	68502	36581	9621	12194	10106
38	0.998	0.722	0.722	76	36	49434	19070	112	68504	36581	9621	12194	10108
37	1.000	0.721	0.721	17	9	49461	19129	26	68590	36581	9621	12194	10194
36	1.000	0.721	0.721	17	9	49461	19129	26	68590	36581	9621	12194	10194
35	1.000	0.721	0.721	17	9	49461	19129	26	68590	36581	9621	12194	10194
34	1.000	0.721	0.721	15	8	49462	19131	23	68593	36581	9621	12194	10197
33	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
32	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
31	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
30	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
29	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
28	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
27	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
26	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
25	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
24	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
23	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
22	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
21	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
20	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
19	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
18	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
17	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
16	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
15	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
14	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
13	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
12	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
11	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
10	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
9	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
8	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
7	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
6	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
5	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
4	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
3	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
2	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220
1	1.000	0.721	0.721	0	0	49470	19146	0	68616	36581	9621	12194	10220

983

1085

1086

AIR FORCE HUMAN RESOURCES LABORATORY

VOCATIONAL INTEREST ASSESSMENT IN THE AIR FORCE

By

William E. Alley

PERSONNEL RESEARCH DIVISION
Lackland Air Force Base, Texas 78236

Paper presented at the 1976 Annual Convention of the Military Testing Association, Pensacola Florida.

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

984 1087

VOCATIONAL INTEREST ASSESSMENT IN THE AIR FORCE

William E. Alley
Personnel Research Division
Air Force Human Resources Laboratory

I. Introduction

The choice of an occupation which later proves to be unsuitable can be extremely costly for the individual making the choice and for his employer. A personal decision to change jobs after considerable preparation can mean the loss of valuable experience and tenure which may be completely unrelated to a new pursuit. The employer, on the other hand, will have invested considerable training resources which cannot be recovered. If the individual persists on the job, the working environment will eventually become a source of chronic dissatisfaction and frustration. The employer will note these effects in the fact that a dissatisfied worker will be less likely to reach full potential as a productive member of the organization.

The problem of vocational choice in the Air Force is complicated by several factors. To begin with, recruits may find it difficult to assess their own personal interests in terms which are meaningfully related to the process of choosing an occupation. They are typically characterized as having little experience in the civilian job market and even less experience with the complex array of Air Force job opportunities. It is small wonder that each year, thousands of recruits enter occupations and later elect to retrain into other specialties at the earliest possible

opportunity. Others continue on in unsuitable work environments and become part of that 10 percent of the work force which purportedly causes 90 percent of the personnel problems.

Vocational interest tests have, for many years, been used to provide information on the appropriate match of people to occupations. The purpose of these inventories is to guide young people into careers which offer the maximum opportunities of obtaining personal satisfaction in their work. How well they accomplish these goals has been subject of considerable debate. Empirical studies of sufficient breadth and scope to evaluate the utility of these measures have been few and contradictory, at least as far as Air Force personnel managers were concerned. Heretofore, aptitude assessment has served as the principal means of selection and classification. Beginning in 1971, coincidental with the move toward an all volunteer recruiting environment, a study was commissioned to evaluate the feasibility of using measured interests as a vocational guidance tool. The principal objectives of this research were:

a. To develop standardized assessment techniques for measuring vocational interests at time of entry into service, and

b. To evaluate the relationships between these measures and later job satisfaction criteria.

Early work on this project performed under Air Force contract (Echternaught, et. al., 1973) resulted in the construction of a 400-item general-purpose vocational interest inventory, designated the Vocational Interest Career Examination (VOICE), specifically for use by the Air Force. This paper describes subsequent in-house efforts to refine the instrument for operational use in guiding recruits toward compatible

occupations in the service. In the first phase of this work, homogeneous subscales were developed to characterize the vocational interests of both male and female recruits. The second phase addressed the need for an empirical basis for using interest measures in the selection of an appropriate occupation. Longitudinal validation studies were conducted to relate interests measured at time of entry with later satisfaction on the job.

II. HOMOGENEOUS SUBSCALE DEVELOPMENT

General purpose vocational interest inventories typically elicit information in the form of relative preferences for various activities, hobbies, occupational types, etc. For purposes of interpreting and using these data for vocational guidance counseling, consideration of individual item responses is extremely inefficient. One approach to this problem is to cluster similar items into relatively homogeneous and separate subscales based on the extent of common variation among the items. Basic to this approach are factor analytic concepts of reducing the dimensional space represented by the original items into a simpler, more interpretable structure. This section describes the development of homogeneous interest keys for the VOICE.

Item Pool

The basic data for these analyses consisted of individual responses to each of the 400 items in the inventory. The VOICE items are categorized into

four a priori areas: jobs, work tasks, spare time activities, and desired learning experiences. The respondent indicates relative preferences for these activities in a standard like/indifferent/dislike format.

Subjects

The inventory was administered to random samples of male (N=10,035) and female (N=12,710) recruits during 1972-1973 at Lackland Air Force Base, Texas. Subjects ranged in age from 18 to 26 years, most had completed high school or some college. In general, they were representative of all volunteer recruits entering the Air Force during that time period. Female recruits, which currently represent approximately 13 percent of accessions, were deliberately oversampled to obtain sufficient numbers of responses from this group.

Factor Analyses

To establish the factor structure of the inventory, item intercorrelation matrices (400 x 400) were obtained separately for the male and female samples. Item responses were coded 3 = like, 2 = indifferent, and 1 = dislike. Each of the matrices was subjected to principal axis factor analysis with varimax rotation (Veldman, 1967). Various extraction limits were examined to determine the appropriate number and character of the resulting factors. Limits were set to provide 16, 21, 26 and 35 varimax factors.

The results of these analyses were virtually identical for the male and female subgroups at each extraction level. Moreover, the factors extracted at lower limits within each sample appeared to be invariant as more and more factors were extracted and rotated. In the process, 18

orthogonal factors were identified and labeled as shown in Table 1. These factors represented 48 percent of the original variance in the male sample and 46 percent in the female sample. The most consistent factors, labeled I-XII in the table, could be seen in the 16-, 21-, 26-, and 35-factor solutions. The remaining dimensions were most identifiable in the 26-factor solution. Representative items for each factor are shown in the table together with the factor loadings attained in the male and female subsamples. The order in which the factors are shown represents the degree to which each factor encompassed smaller and smaller item clusters. The office administration factor, for example, had more than 60 items in the inventory with associated loadings above .30. The automated data processing factor had approximately seven items above this minimum cutoff.

Unit-Weighted Subscales

The development of unit-weighted subscales corresponding to these factors followed more or less traditional guidelines. For each factor, items were selected which had factor loadings greater than .30 up to a maximum of 20 items. The unit-weighted subscales have the advantage of simplified scoring and of reducing reliance on all 400 items to generate a single factor score. To verify that the unit weighted subscales did in fact replicate the factors as defined in the analysis, the matrix of these scores was used to predict each of the factor scores in turn. The results of these analyses, shown in Table 2, suggested that the factors and corresponding unit-weighted scales were highly correlated and that linear combinations of the unit-weighted scores replicated the original factors almost entirely.

Normative characteristics of the unit-weighted VOICE subscales are shown in Table 3. It will be noted that the internal consistency of these

Table 1. Orthogonal Interest Factors

Factor/Representative Items	Factor Loading Male	Female
I Office Administration		
Keep personnel records on employees	.74	.70
Organize a file system for an office	.72	.73
Take inventory for a department store	.70	.66
II Electronics		
Find a problem in an electric circuit and fix it	.79	.80
Test television tubes	.72	.73
Technician (electronics)	.72	.69
III Heavy Construction		
Pour concrete for highway construction	.69	.61
Operate a bulldozer or power shovel	.64	.58
Construction worker	.62	.55
IV Science		
Write a scientific report	.65	.68
Work in a scientific laboratory	.79	.82
Use chemical laboratory apparatus	.75	.77
V Outdoors		
Go canoeing	.60	.58
Learn survival techniques	.61	.48
Spend a week at the seashore	.67	.61
VI Medical Service		
Perform physical therapy	.59	.62
Assist a surgeon during an operation	.60	.73
Dental hygienist	.55	.57
VII Aesthetics		
See a Broadway play	.48	.50
Read Shakespeare's plays	.59	.57
Listen to an opera	.61	.66

Table 1 . (Continued)

Factor/Representative Items	Factor Loading Male	Female
VIII Mechanics		
Tune-up a car	.66	.51
Adjust the brakes on an automobile	.61	.46
Supervise work in a garage	.44	.29
IX Food Service		
Work as a short-order cook	.62	.62
Manage a cafeteria	.45	.54
Chef	.65	.70
X Law Enforcement		
Highway patrolman	.71	.68
Private investigator	.56	.52
Prison guard	.56	.52
XI Audiographics		
Photographer	.65	.69
Operate a 16mm movie camera	.62	.67
Record the sound track for a motion picture	.48	.55
XII Mathematics		
Solve arithmetic problems	.67	.75
Algebra	.61	.67
Devise short-cut methods for adding numbers	.48	.52
XIII Agriculture		
Farmer	.54	.57
Work as a game warden	.47	.55
Train animals	.41	.54
XIV Teacher/Counseling		
Listen to people's problems and try to help them	.40	.53
Give on-the-job training	.35	.48
Teacher	.36	.49

Table 1 . (Continued)

Factor/Representative Items		Factor Loading Male	Female
XV	Marksman		
	Gunsmith	.55	.55
	Go deer hunting	.48	.53
	Teach marksmanship	.48	.46
XVI	Craftsman		
	Tailor	.35	.56
	Jeweler	.30	.30
	Watchmaker	.34	.24
XVII	Drafting		
	Draw graphs	.52	.50
	Draw maps from photographs	.48	.49
	Draw a topographical map of the US	.46	.46
XVIII	Automated Data Processing		
	Computer operator	.70	.77
	Computer programmer	.70	.77
	Keypunch operator	.45	.53

Table 2 . Relationships Between VOICE Subscales[a] and Orthogonal Interest Factors

Factor		VOICE SUBSCALES	
		Corresponding Scale (R)	All Scales Combined (Mult R)
1	Office Administration	94[b]	97
2	Electronics	94	98
3	Heavy Construction	88	96
4	Science	87	98
5	Outdoors	83	93
6	Medical Service	79	97
7	Aesthetics	73	93
8	Mechanics	73	95
9	Food Service	79	96
10	Law Enforcement	84	95
11	Audiographics	72	93
12	Mathematics	71	94
13	Agriculture	69	92
14	Teacher/Counseling	52	80
15	Marksman	69	86
16	Craftsman	49	71
17	Drafting	69	90
18	Automated Data Processing	65	90

[a] Unit Weighted

[b] Decimals Omitted

TABLE 3. VOICE Subscales

Scale	No of Items	Score Range[a]	MALES (N=10035) $\bar{x}$	MALES (N=10035) SD	MALES (N=10035) α[b]	FEMALES (N=12710) $\bar{x}$	FEMALES (N=12710) SD	FEMALES (N=12710) α[b]
Office Administration (OA)	20	20-60	32.32	10.71	.95	37.85	11.46	.95
Electronics (EL)	20	20-60	40.72	12.78	.96	32.50	12.15	.96
Heavy Construction (HC)	20	20-60	34.41	9.99	.93	27.49	8.37	.93
Science (SC)	20	20-60	38.02	12.69	.96	38.21	12.81	.96
Outdoors (OD)	15	15-45	36.47	6.88	.88	36.67	5.76	.83
Medical Service (MS)	20	20-60	33.38	10.55	.94	40.81	11.47	.94
Aesthetics (AE)	15	15-45	26.10	7.74	.90	31.78	7.36	.88
Mechanics (ME)	15	15-45	31.65	8.98	.94	25.10	8.69	.94
Food Service (FS)	15	15-45	21.36	6.30	.90	26.72	7.40	.90
Law Enforcement (LE)	15	15-45	29.22	7.30	.88	26.90	6.89	.86
Audiographics (AU)	10	10-30	20.76	5.79	.90	22.26	5.45	.88
Mathematics (MA)	12	12-36	21.50	7.37	.93	22.06	7.44	.92
Agriculture (AG)	15	15-45	28.00	7.15	.88	31.04	8.00	.90
Teacher/Counseling (TC)	10	10-30	19.32	5.73	.89	22.22	5.31	.86
Marksman (MK)	7	7-21	15.38	4.28	.86	11.54	4.22	.86
Craftsman (CF)	7	7-21	9.88	2.95	.79	11.22	2.96	.72
Drafting (DF)	7	7-21	13.25	4.20	.85	13.08	4.28	.86
Automated Data Processing (DP)	7	7-21	13.76	4.49	.89	13.86	4.39	.88

[a]Items scored 3 = Like; 2 = Indifferent; and 1 = Dislike; missing or otherwise invalid responses recoded = 2.

[b]Alpha coefficient of internal consistency (Cronbach, 1951).

scales was quite high, ranging from .72 to .96. Male and female differences in means and standard deviations, also shown in the table, indicate that, males typically score higher on the Electronics, Heavy Construction, Mechanics, Law Enforcement, and Marksman subscales. Female groups typically scored higher on Office Administration, Aesthetics, Medical Service, Food Service, Audiographics, and Teacher/Counseling as might be expected.

III. PREDICTIVE VALIDATION STUDIES

To be of practical value in counseling young recruits, vocational interest inventories must provide information relevant to the selection of the most appropriate occupation. That is, these measures must elicit information from occupationally naive respondents which can then be used to estimate eventual satisfaction in a wide variety of alternative careers. The predictive validity of the VOICE subscales was examined by administering follow-up surveys after one year on the job to determine the degree of satisfaction with assignment. The extent to which vocational interests, measured prior to entry into service, could be used to predict satisfaction within various specialties served as the basis for developing empirical scales applicable to specific occupational groups, and for evaluating the overall accuracy of the system.

Procedure

Job satisfaction surveys were mailed to 18,000 recruits who had met the criteria for job tenure. The purpose of the survey was to determine the Air Force occupation to which they were assigned and how well satisfied they were with the assignment. Occupations in the Air Force are identified by five-digit specialty codes (AFSC's) which distinguish between career fields

and between various skilled levels within fields. For purposes of this study, the AFSC's were grouped into 20 job clusters as defined in the DoD occupational conversion table (U.S. Department of Defense, 1975). A description of these clusters is shown in Table 4.

Satisfaction with assigned AFSC was obtained from responses to the following survey item: How satisfied are you with your present AFSC? In the analysis, responses to this item were coded 800 = very satisfied, 600 = moderately satisfied, 400 = moderately dissatisfied, and 200 = very dissatisfied. The return rate in the follow-up survey was 69 percent overall. Excluding unusable returns, there were 7,521 valid cases in the male sample and 3,767 in the female sample (N = 11,288). The distribution of cases across the 20 job groups is also shown in the table.

In addition to the vocational interest data available for each subject, percentile scores from the Armed Services Vocational Aptitude Battery (ASVAB) were obtained from computerized record files maintained at the laboratory. The ASVAB is a differential aptitude test used by the Department of Defense (DoD) for selection and classification purposes. The battery takes two hours to complete and yields four standard aptitude indexes (Administrative, Electronics, General, and Mechanical). The aptitude data was included in the analysis for statistical control purposes. Since this information was typically available during standard career counseling interviews, a portion of the analysis was devoted toward establishing any relationships which might exist between job satisfaction and individual aptitudes.

Analyses

To investigate the effects of pre-service interest and aptitude measures on later job satisfaction, multiple regression analyses (Ward & Jennings, 1973)

Table 4. DoD Occupational Groups and Sample Distribution

Occupational Group		Sample		
Code	Description	Male	Female	Total
10	Radio/Radar Equipment Repair	472	175	647
1X	Misc Electronic Equipment Repair	466	165	631
22	Radar and Air Traffic Control	227	76	303
2X	Misc Communications and Intelligence Specialties	241	148	389
30	Medical Care	257	226	483
3X	Misc Medical and Dental Specialties	121	86	207
4X	Technical and Allied Specialties	169	103	272
51	Administration	855	922	1777
5X	Misc Administrative Specialties and Clerks	596	530	1126
600	General Aircraft Mechanic	886	480	1366
601	Aircraft Engine Mechanic	294	117	411
602	Aircraft Accessories Mechanic	417	178	595
64	Armaments and Munitions	415	-	415
6X	General Mechanic	296	69	365
72	Utilities and Maintenance	134	43	177
78	Firefighters	162	-	162
82	Material Receipt, Storage and Issue	422	133	555
83a	Security Police	651	-	651
83b	Law Enforcement	202	149	351
8X	Misc. Services and Supply	238	167	405
	TOTAL	7521	3767	11288

were performed within each DoD cluster. In this procedure, the errors of prediction associated with using a given set of measures (full model) were compared with the errors associated with a reduced set (restricted model). The restricted models are defined so as to represent the test of a specific null hypothesis for which F-statistics and associated probability values can be computed. Since the number of these comparisons within each occupational cluster was quite large, selected system-wide F-tests were first conducted to provide control over the Type I error rate. Using available information on all valid cases, the following sequence of questions was addressed:

a. Are there any differences in reported satisfaction attributable to DoD group membership, sex category, VOICE, or ASVAB variables separately weighted for each sex and occupational group?

b. Are there any differences in reported satisfaction attributable to ASVAB scores controlling for effects due to sex and DoD group membership?

c. Are there any differences in reported satisfaction attributable to VOICE subscale performance controlling for effects due to sex and DoD group membership?

d. Are there any differences in reported satisfaction attributable to ASVAB scores controlling for effects due to sex, DoD group membership and VOICE subscale performance?

e. Are there any differences in reported satisfaction attributable to VOICE subscale performance controlling for effects due to sex, DoD group membership, and ASVAB performance?

Based on the system-wide results, further analyses were performed within each of the 20 DoD occupational groups to assess the independent contribution of the test variables with and without distinctions based

on sex. Summary documentation was obtained in the form of zero order correlations of each separate predictor variable with the satisfaction criterion, multiple correlations, F-ratios, and associated probabilities.

Job Satisfaction

Subsample means and standard deviations on the satisfaction criterion for each of the 20 specialties are shown in Table 5. An inspection of these values shows the highest degree of satisfaction with assignment among male respondents was reported in electronics repair (10), medical technician (30 and 3X), technical and allied specialties (4X), utilities maintenance (72), and the fire fighting career field (78). Least satisfied were respondents from the armaments and munitions career field (64), materiel receipt, storage, and issue (82), and security police (83a). In the female sample, somewhat less satisfaction was noted in electronics repair (10) and considerably less in the mechanical (600 thru 6X) and utilities maintenance (72) career fields. Overall, these results correspond very closely to those found in previous studies of Air Force job satisfaction (Alley & Gould, 1975; Gould, 1972; Echternaught, et. al., 1973) lending some credibility to the criterion assessment.

Effects of Aptitude and Interests

Analyses to determine the effects of entry level interests and aptitudes on later satisfaction with assignment are summarized below. Results of system-wide comparisons involving the total sample are shown in Table 6. Evidence from the first comparison indicates a significant overall effect due to one or more of the predictor sets. Significant baseline effects due to simple group membership were documented in the second comparison. That is, general satisfaction with assignment was found to differ among the sex and

Table 5. Criterion Means and Standard Deviations by Sex and DoD Occupational Group[a]

Occupational Group Code	Description	Male $\bar{X}$	Male SD	Female $\bar{X}$	Female SD
10	Radio/Radar Equipment Repair	640	147	584	181
1X	Misc Electronic Equipment Repair	588	181	587	187
22	Radar and Air Traffic Control	611	213	579	224
2X	Misc Communications and Intelligence Spec	561	197	520	201
30	Medical Care	644	160	668	162
3X	Misc Medical and Dental Specialties	693	139	705	139
4X	Technical and Allied Spec	644	188	660	160
51	Administration	556	190	583	190
5X	Misc Administrative Specialties & Clerks	593	187	582	183
600	General Aircraft Mechanic	548	191	491	223
601	Aircraft Engine Mechanic	615	171	523	217
602	Aircraft Accessories Mechanic	526	184	465	219
64	Armaments and Munitions	434	198	-	-
6X	General Mechanic	583	197	516	188
72	Utilities Maintenance	645	179	530	206
78	Firefighters	653	188	-	-
82	Material Receipt, Storage and Issue	452	195	484	197
83a	Security Police	440	208	-	-
83b	Law Enforcement	610	197	605	199
8X	Misc Services and Supply	524	198	472	210

[a]Job Satisfaction Scale:

800 = Very Satisfied
600 = Moderately Satisfied
400 = Moderately Dissatisfied
200 = Very Dissatisfied

Table 6. Results of System-Wide Regression Analyses

	Models		R^2		d_{f1}	d_{f2}	F
	Full	Rest	Full	Rest			
(1)	(VOICE + ASVAB) x Group	Unit Vector	.2141	.0000	850	10437	3.35**
(2)	Group	Unit Vector	.1068	.0000	36	11251	37.37**
(3)	ASVAB x Group	Group	.1201	.1068	148	11103	1.13^{ns}
(4)	VOICE x Group	Group	.2027	.1068	666	10585	1.91**
(5)	(VOICE + ASVAB) x Group	VOICE x Group	.2141	.2027	148	10437	1.02^{ns}
(6)	(VOICE + ASVAB) x Group	ASVAB x Group	.2141	.1201	666	10437	1.87**

** Significant beyond the .01 level

ns Non-significant

DoD occupational groups. The additional contribution of aptitude measures to the baseline variables was evaluated in the third and fifth comparisons and was found to be non-significant. In comparison 3, the aptitude indices weighted separately for each sex and occupational group did not improve prediction over and above the accuracy obtained using group membership variables alone. Nor did they contribute when both group membership and interests were held constant in comparison 5.

Vocational interests, on the other hand, were found to make significant and unique contributions to prediction when both group membership and aptitudes were held constant (comparisons 4 and 6). Based on these results, it was concluded that satisfaction with assignment was functionally related to both prior service interests and group membership (sex and DoD occupation). No evidence was found that aptitudes influenced eventual satisfaction with assignment. In all subsequent analyses, the aptitude variables were eliminated from further consideration as a basis for predicting job satisfaction. Emphasis was directed instead at further clarification of the vocational interest effects as modified by occupational group and sex categories.

Results of Analyses Within Occupational Groups

Results of analyses within specialties are shown in Table 7. The first series of comparisons indicates the unique effects associated with sex. These models tested the hypothesis that all the coefficients in the male equations were equal to corresponding coefficients in the female equations within each specialty. As can be noted in in the table, no unique sex differences in the VOICE equations were detected in 13 of the 17 career

Table 7. Summary of Regression Analyses Within Occupational Groups

	Occupational Group	Multiple Correlation			Significance of Unique Contribution	
		Model 1: VOICE x Sex	Model 2: VOICE	Model 3: Sex	Sex	VOICE
10	Radio/Radar Equip Repairman	.33	.27	.16	ns	**
1X	Misc Electronic Equip Repair	.32	.26	.00	ns	**
22	Radar and Air Traffic Contrl	.42	.32	.06	ns	*
2X	Misc Comm and Intell Spec	.32	.26	.10	ns	ns
30	Medical Care	.36	.29	.07	ns	**
3X	Misc Medical and Dental Spec	.47	.42	.04	ns	**
4X	Technical and Allied Spec	.28	.22	.04	ns	ns
51	Administration	.32	.28	.07	**	**
5X	Misc Admin Spec and Clerks	.26	.23	.03	ns	**
600	Gen'l Aircraft Mechanic	.32	.28	.13	*	**
601	Aircraft Engine Mechanic	.48	.41	.22	*	**
602	Aircraft Accessories Mechanic	.39	.32	.14	ns	**
64	Armaments and Munitions	-	.21	-	-	ns
6X	General Mechanic	.37	.31	.14	ns	**
72	Utilities Maintenance	.54	.35	.26	*	ns
78	Firefighters	-	.37	-	-	ns
82	Material Receipt, Storage & Issue	.27	.23	.07	ns	*
83a	Security Police	-	.35	-	-	**
83b	Law Enforcement	.44	.36	.01	ns	**
8X	Misc Services and Supply	.38	.32	.12	ns	**

* Significant at the .05 level
** Significant at the .01 level
ns Non-significant
-) Denotes occupations restricted to male entrants

fields containing members of both sex. In these specialties, a common VOICE equation without sex distinctions yielded equivalent accuracy to that obtained with separate equations for both males and females. Exceptions to this trend were noted in the administrative (51), mechanical (600, 601,) and utilities maintenance (72) career fields where the male and female equations were found to differ significantly.

The next series of comparisons was made to determine if the VOICE weights were significantly different from zero, holding the sex variable constant. In 15 of the 20 comparisons, significant prediction systems using the VOICE subscales were detected. Those career fields in which satisfaction was not found to be related to pre service interests included the 2X, 4X 64, 72, and 78 career fields. In all other cases, moderate but highly significant relationships were found to exist between interests as measured by the VOICE subscales and later satisfaction on the job. Viewed in terms of the number of personnel included in the study, the non-significant career fields included roughly 13 percent of the total sample while 87 percent fell within significant prediction systems.

Insight into specific relationships between interests and satisfaction within each specialty may be found in Table 8 which summarizes the most significant correlates among the VOICE subscales. In reviewing these findings, it is important to note that the table catalogues zero order relationships between each subscale viewed as an independent measure and the satisfaction criteria, rather than net relationships, which may show a somewhat difficult pattern.

In the electronics field (10 & 1X), satisfaction was positively correlated with entry-level scores on the Electronics, Heavy Construction,

Table 8. VOICE Subscale Correlates of Job Satisfaction in Selected Career Areas

Positive	Negative
Electronics (10 & 1X)	
Electronics	Aesthetics
Heavy Construction	Medical Service
Mechanics	
Law Enforcement	
Marksman	
Automated Data Processing	
Radar and Air Traffic Control (22)	
Science	
Outdoors	
Drafting	
Medical Technician (30 & 3X)	
Science	Audiographics
Medical Service	
Food Service	
Agriculture	
Administration (51 & 5X)	
Office Administration	Augiographics
Automated Dat Processing	Marksman
	Drafting
Mechanics (600, 601, 602 & 6X)	
Electronics	Office Administration
Heavy Construction	Medical Service
Outdoors	Food Service
Mechanics	Craftsman
Law Enforcement	
Agriculture	
Marksman	
Material Receipt, Storage and Issue (82)	
Office Administration	
Heavy Construction	
Food Service	

Table 8. (Continued)

Positive	Negative
Police (83a & 83b)	
Law Enforcement	Electronics
Agriculture	Science
Marksman	Medical Service
	Audiographics
	Mathematics
	Automated Data Processing
Services & Supply (8X)	
Heavy Construction	Medical Service.
Food Service	
Law Enforcement	
Agriculture	

Mechanics, Law Enforcement, Marksman, and Automated Data Processing subscales. On the other hand, dissatisfaction in this career was associated with high scores on Aesthetics and Medical Service. Persons assigned to Radar and Air Traffic Control (22) reported satisfaction in direct proportion to their scores on the Science, Outdoors, and Drafting subscales. Satisfied medical technicians (30 & 3X) were more likely to have high scores in Science, Medical Service, Food Service, and Agriculture while obtaining relatively low scores in Audiographics. In the Administrative careers (51 & 5X), positive relationships were found between eventual satisfaction and interests in Office Administration and Automated Data Processing. For the most part, relationships between interests and job satisfaction noted these and the remaining careers followed a more or less predictable pattern. In occupations where there was a logical counterpart among the VOICE subscales, positive correlations were usually detected. Entry level scores on the Mechanics subscale, for example, were predictive of satisfaction in the mechanical career (600, 601, 602 & 6X) while scores on the Law Enforcement and Marksmanship scale provided useful indicators of eventual satisfaction in police work (83a & 83b).

Overall, these results were very encouraging from a guidance standpoint, inasmuch as the capability now exists for measuring interests among new recruits and relating this information directly to the appropriate choice of a career.

IV. APPLICATIONS

The application of this methodology to a specific respondent requires at least two vectors of information as shown in the first two columns of

Table 9. In the score vector are each of the 18 subscale scores obtained on the VOICE. In the regression weight vector are the appropriate weights for estimating satisfaction in a specific career area--in this case, the medical career field (30). This recruit's expected satisfaction in the medical area, it is estimated by cross-multiplying the raw scores with the appropriate regression weights for each subscale and adding the products (Column C) across all scales. The result is adjusted by the constant shown at the bottom of the figure. The man's estimated satisfaction is found to be 675 out of a possible 800 points. For comparison purposes, the score of a person at the average on each of the VOICE subscales is computed to be 606 out of a possible 800. Thus, on the basis of his interests at time of entry into service, he would be more likely to be satisfied in the medical field (30) after a year on the job than would the average recruit.

Expected values for this respondent computed for each of the 20 occupations used in the previous analysis are shown graphically in Figure 1 (dotted line). For reference purposes, the expected values of an average recruit have also been plotted (solid line). In terms of absolute satisfaction, this man would probably be best suited for medical care (30) or the miscellaneous medical and dental specialties (3X). Least optimal from the standpoint of expected satisfaction would be assignments to security police (83a) or the armament and munitions (64) career field. Figure 2 shows a corresponding profile for a female recruit. Since female entry into three of the twenty groups was restricted at the time of the study, seventeen estimates are available. The large negative depression across the mechanical careers (600, 601, 602 & 6X) indicates this recruit would probably

Table 9. Estimating Satisfaction in the Medical Care (30) Specialties

Subscales	Raw Score	Raw Score Weights for 30 Medical Care	Products (1) x (2)
Office Administration	58	2.18	126.44
Electronics	60	.21	12.60
Heavy Construction	42	.78	32.76
Science	60	1.60	96.00
Outdoors	45	- .99	- 44.55
Medical Services	48	3.20	153.60
Aesthetics	45	- .07	- 3.15
Mechanics	43	.17	7.31
Food Service	35	1.94	67.90
Law Enforcement	37	.24	8.88
Audiographics	28	.61	17.08
Mathematics	36	-1.14	- 41.04
Agriculture	43	.99	42.57
Teacher/Counseling	30	-3.73	-111.90
Marksman	21	1.03	21.63
Craftsman	15	-4.44	- 66.60
Drafting	21	-5.98	-125.58
Automated Data Processing	14	-2.28	- 31.92
	Constant	512.90	512.90
			$\sum$ = 674.93

Figure 1. VOICE Occupational Profile -- Male

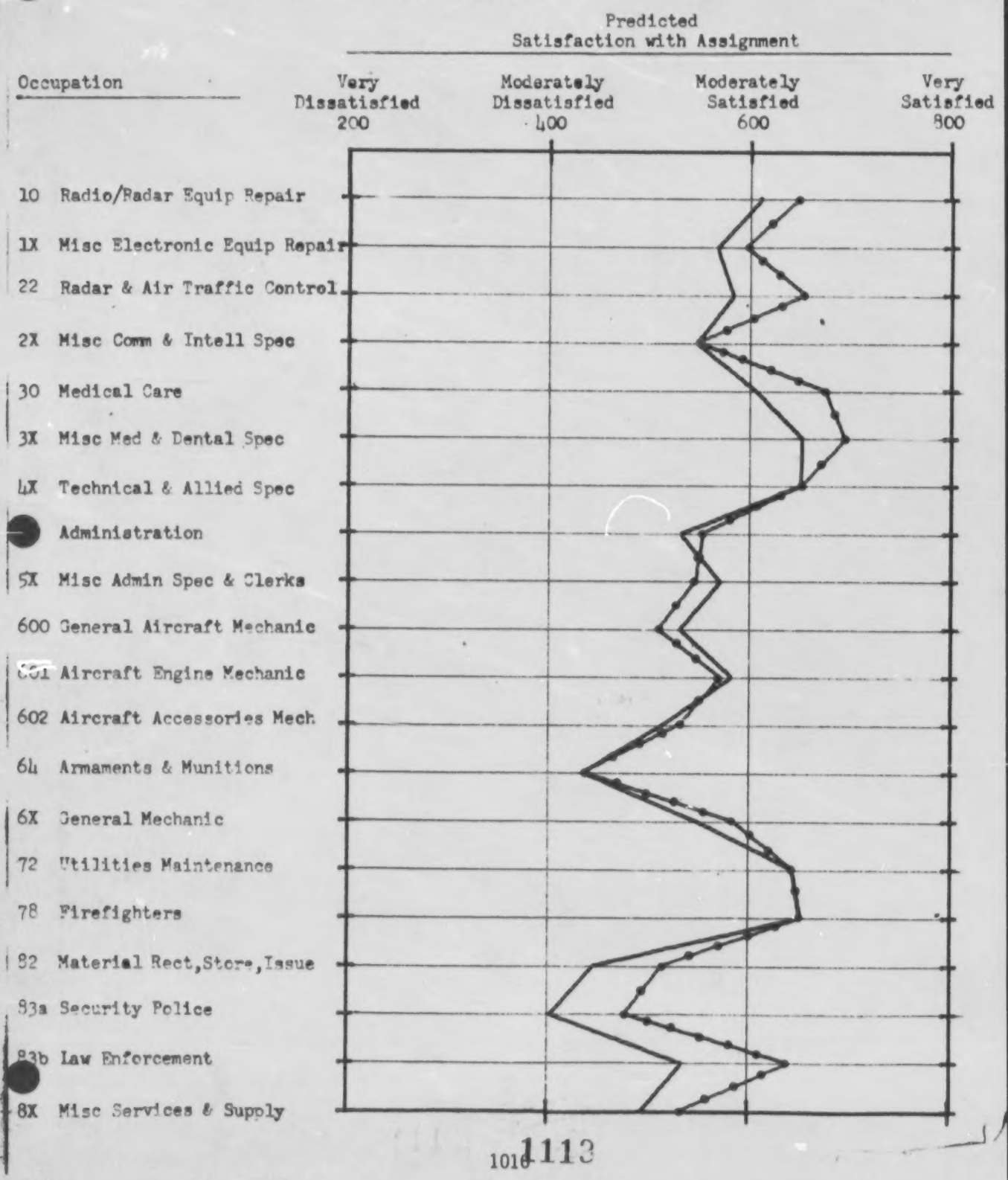
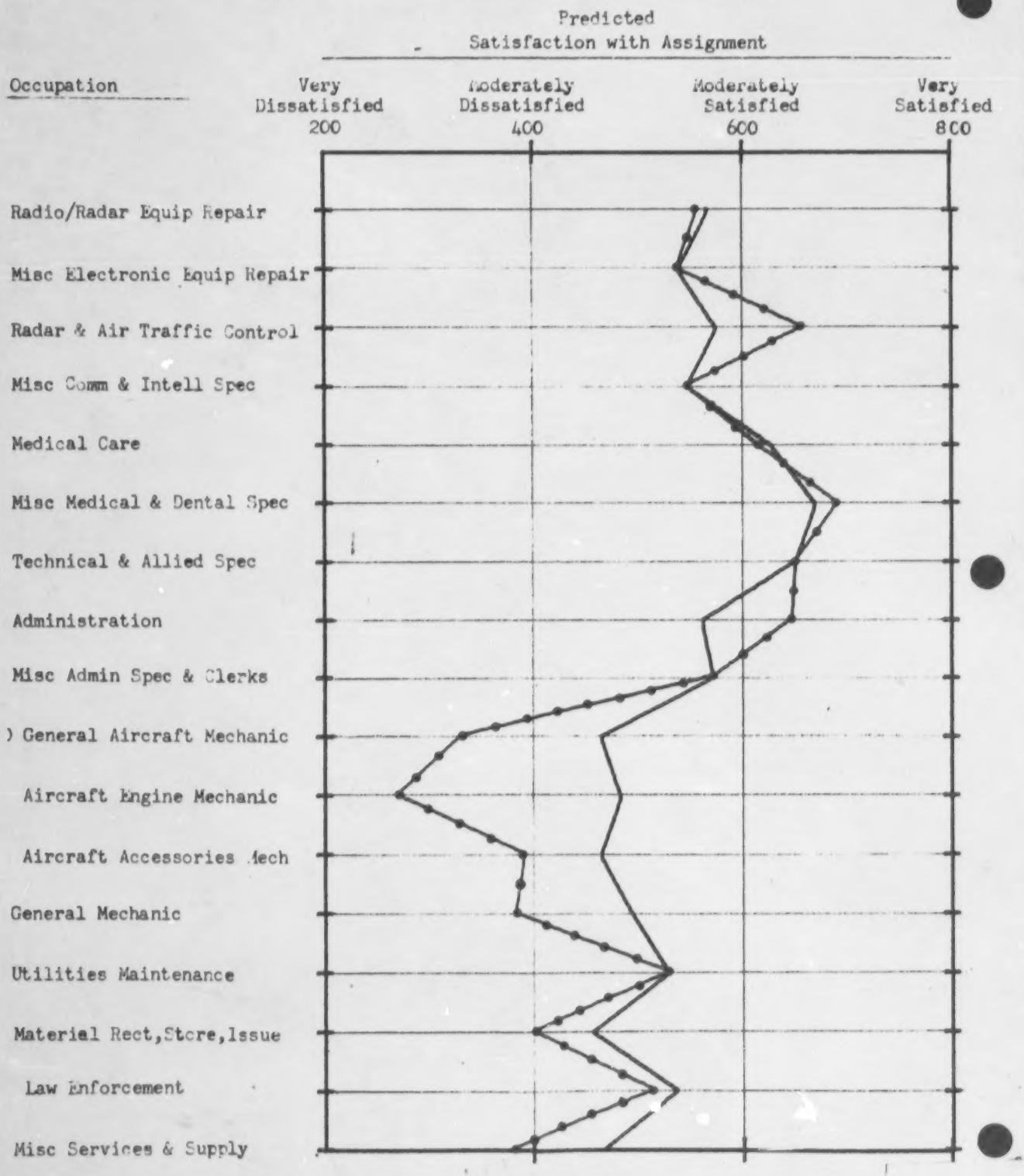


Figure 2. VOICE Occupational Profile -- Female



be ill-suited for that area of assignment. On the other hand, there are indications that she might be more appropriately assigned to miscellaneous medical and dental field, radar/ATC, technical and allied specialties or possibly administration.

V. CONCLUSIONS AND RECOMMENDATIONS

The principal research findings of this study can be summarized as follows:

a. Vocational interests, as measured by the VOICE can be characterized by relatively few independent measurements that, in form and substance, are virtually identical for male and female respondents. In the present study, eighteen common interest dimensions, identified using factor-analytic techniques, accounted for 48 percent of the item variance in the male sample and 46 percent in the female sample.

b. A simple unit-weighted scoring technique for these dimensions replicated the original factors almost entirely. Internal consistency of items within the homogeneous subscales ranged from .79 to .96 among males and from .72 to .96 among females.

c. Normative data based on 10,035 male respondents and 12,710 female respondents indicate that, while the scales have identical meaning for both sexes, the degree of preference associated with each was not always similar. Males typically scored higher on Electronics, Heavy Construction, Mechanics, Law Enforcement, and Marksman, while females as a group had higher mean scores on Office Administration, Medical Service Aesthetics, Food Service, Agriculture, and Teacher/Counseling. Sex differences on the remaining scales were negligible.

d. In the longitudinal validation study, moderate but highly consistent relationships were found to exist between vocational interests measured at the point of entry into service and reported satisfaction with assignment after approximately one year on the job.

e. These relationships were statistically significant in and of themselves and when baseline effects due to aptitude and simple group membership were held constant. The pattern of relationships between specific interests and satisfaction appeared to vary considerably by occupational group and within groups, occasionally varied by sex category, although this latter finding was true for only a small proportion of the career fields examined. For the most part, the regression equations relating interest to satisfaction were homogeneous with respect to sex.

f. No evidence was found that aptitude scores as measured by the ASVAB were related to reported satisfaction.

Future research efforts in this area are being directed at (a) increasing the sample of respondents to provide more definitive classification guidelines, (b) further exploration of commonalities between career fields, (c) norming and standardization of operational composites on a nationwide sample of high school students, and (d) investigation of other criteria of interest which might be effected in a positive way with more sophisticated utilization of vocational interest data (i.e., technical training and job tenure). Additional documentation for the work described in this report may be found in Alley, Wilbourn, & Berberich, (1976) and Alley, Berberich and Wilbourn (1976).

References

Alley, W. E., Berberich, G. L. & Wilbourn, J. M. Dimensions of vocational interest. AFHRL-TR-76-00. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, 1976 (in press).

Alley, W. E., Wilbourn, J. M. & Berberich, G. L. Relationships between performance on the Vocational Interest Career Examination and reported job satisfaction. AFHRL-TR-76-00. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, 1976 (in press).

Alley, W. E. & Gould, R. B. Feasibility of estimating personnel turnover from survey data - A longitudinal study. AFHRL-TR-75-54. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, October 1975.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrica, 1951, 16, 297-334.

Echternacht, G. J., Reilly, R. R., & McCaffrey, P. J. Development and validity of a vocational and occupational interest inventory. AFHRL-TR-73-38. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1973.

Gould, R. B. Reported job interest and perceived utilization of talents and training by airmen in 97 career ladders. AFHRL-TR-72-7. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, January 1972.

United States Department of Defense, Office of the Assistant Secretary of Defense (M&RA). Occupational Conversion Table. DoD 1312.1-E. Washington, D. C.: U.S. Government Printing Office, 1975.

Ward, J. H., & Jennings, E. Introduction to linear models. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Veldman, D. J. Fortran programming for the behavioral sciences. New York: Holt, Rinehart and Winston, 1967.

EVALUATION OF FIVE FOOD-RATING-FEEDBACK CARDS

Herbert Meiselman, Day Waterman, Nancy Cobean

Food Sciences Laboratory

US Army Natick Research & Development Command

Natick, Massachusetts 01760

EVALUATION OF FIVE FOOD-RATING-FEEDBACK CARDS[1]

The purpose of this report is to present an evaluation of a food acceptance rating card for use in military dining halls. In a series of studies of consumer's opinions of military food service consumers have stated that their most serious complaint concerns food quality (Jacobs and Meiselman, 1976)[2]. Thus, the key issue in improving military food is to improve the food. This requires food service system planners and managers to be knowledgeable about the quality of their product. There are two general approaches to obtaining this type of information: one is to utilize professional or semiprofessional staff to inspect food; another is to depend on feedback from customers about the food. A combination of approaches is probably the best answer. Ideally, feedback from customers should be obtained in an economical manner, and in a manner which avoids possible bias on the part of those collecting the data. Circumventing the dining facility management in the collection process would help to insure the elimination of bias. A feedback card which can be self-administered by customers is ideal from the practical viewpoint. Therefore, if it can be shown to be a good measure of

---

[1] This report is based on US Army Natick R&D Command Technical Report 76-58-FSL.

[2] Jacobs, H. L. and H. L. Meiselman. Customer morale and behavioral effectiveness: Accomplishments and goals of psychological studies of food service systems. In: Proceedings, Third International Meeting, Foods for the Armed Forces. NARADCOM Technical Report 76-42-OTD, March, 1976.

opinion, the feedback card could have an important place in military food service.

Members of the Behavioral Sciences Division, Food Sciences Laboratory of the U.S. Army Natick Research & Development Command (NARADCOM) were approached informally by personnel of the Food Service Directorate, U.S. Army Troop Support Agency (USATSA) to aid in evaluating a new consumer food preference rating card designed by USATSA. This card (Appendix A, Figure A-1) utilizes 3 behavior/attitude questions followed by eight rating scales. The original request entailed analysis of rating card data which had been collected by USATSA in dining facilities at Fort Lee. These data included eight different food items. Following initial testing, data from three foods (grilled steak, mashed potatoes, and cereal) were selected for a more thorough evaluation. These foods were chosen for two reasons: they had relatively large numbers of subjects responding, and they represented the spectrum of foods tested.

Following the initial analysis of USATSA data, it was determined that NARADCOM should develop several alternative forms of preference rating cards, the performance of which could be compared with that of the USATSA card. The testing of these new style cards took place at NARADCOM and Ft. Lee. The key question addressed by the NARADCOM was whether any or all of the card forms could detect intention adulteration of a food (e.g., reduced portion size or different grades of beef) and thereby provide a measure of validity?

Thus, the purpose of this investigation was to study the statistical properties of responses to the feedback card developed by USATSA and

compare these with responses to alternative card forms developed by NARADCOM. A key test was to determine whether any or all cards would reflect intentional adulteration of a food.

METHOD

The major portion of this investigation centers around the collection of food acceptance data using a 3" x 8" or 5" x 8" rating card. Although there were multiple administrations of this survey at different times of the year and day, or at different physical locations, it is possible to describe the general procedure of data collection for all cases at one time.

Respondents: The respondents were primarily enlisted personnel below the rank of E-5. Most were males between the ages of 18 and 21. The schools at Ft. Lee did have some senior NCO's, however.

Location: Surveying took place in Class A dining facilities, both short order and regular menu lines. Several of these facilities were of the consolidated type.

Food Selection and Preparation: The foods surveyed were generally selected on the occasion of their appearance in the 42 day Master Menu. An exception was made in several cases in order to collect data on the same food on consecutive days. The preparation of foods used standard procedures in accordance with the Army Forces Recipe Cards customarily used by all dining facilities. The exception involving white cake will be discussed later.

Procedure: Signs were posted on the outer doors to the dining facility announcing the test food. The "headcount" was requested to hand survey cards to the first one hundred willing participants as the individuals signed in at the desk. In one session the "headcount" was asked to hand-out 200 cards because a large number of respondents was critical for adequate statistical analysis. The food being evaluated (only one per meal) was available on the serving line as it would have been at any other time. No attempt was made to entice the subjects to select it. At times, however, the test food was the only one in its class available. This happened by chance when the alternate choice was depleted, or intentionally when the experimenter wished to better the chances of getting a large response to the food items. At the end of the serving line or near the dish bussing line, a sign identified a receptacle where the participant was to voluntarily return the card after his meal. The instruction on the cards varied in their length and detail (See Appendix A). On several occasions, cards were retrieved from the table tops following the meal. These cards were analyzed separately.

RESULTS AND DISCUSSION

Original USATSA Data

These data, collected in six dining facilities at Fort Lee, evaluated eight food items. The foods which were served at the noon and evening meals were testing using the USATSA-designed rating card (See Appendix A, Figure A-1). Data from three foods

(grilled steak, mashed potatoes, and cereal) were selected for in-depth analyses, including analysis of variance (ANOVA) and discriminant analysis.

The three behavior/attitude questions on the card were the initial point of focus in this investigation. These questions were: "Did you select_____? Do you normally like this item? and Do you like to eat here often?" Although persons who had indicated they had not selected the food item were instructed not to complete the rating card, many did. The ratings given by non-selectors, however, were not statistically different from those given by selectors. It is unclear whether those who stated they had not selected the food: (a) were correct in stating they had not selected it, and (b) were responding to a previous memory of the food in the dining hall at home, or at a commercial facility, or (c) were using information supplied by someone else. Nevertheless, the feedback card must be viewed as a response to an item as it is served on a particular day. The card forms designed later by NARADCOM changed the format to try to eliminate the inappropriate responses. The lengthy instructions on Card #1 were omitted (See Appendix A). The selection question simply read:

Did you select_____? Yes (continue answering)
No (STOP and Return Card)

This question was separated from the rest of the card by a bold line.

The second behavior/attitude question, "Do you normally like this item?" was shown by discriminant analysis to be the best predictor of whether or not the subject selected the item. There was

also a significant difference between the ratings of "likers" and "non-likers". It is not surprising to discover that people who do not normally like the food, reflect that opinion in the ratings of the food, assuming the food and preparation methods are constant. Nonetheless, it is our opinion that anyone who has selected the food (and presumably eaten some or all of it) is entitled to an opinion about it. Therefore, elimination of this question was advised for simplicity's sake.

The discriminant analysis uncovered a strong relationship between the third question, "Do you like to eat here often?", and the eight evaluative questions (good/bad). This relationship indicated that rating of a food was not independent of the degree of like of the dining facility, i.e., people who disliked the dining facility, also disliked the food, possibly independent of the food quality. Since one cannot easily resolve whether people dislike the dining facility because of the food, or the food because of the dining facility, or dislike both because of a third factor, it was decided to delete this question from the form developed by NARADCOM.

The eight ratings of food attributes were found to be significantly correlated (Pearson r). This finding suggests that although the categories are logically dissimilar (e.g., texture vs. portion size), the subjects are not responding to them independently. They are responding with a bias which does not reflect a true multiattribute evaluation of the food. Many hypotheses can be made as to why this occurred. Reducing the number of food qualities to be rated from eight to four, thereby simplifying the task, was expected to encourage the subjects to respond differentially across the food qualities.

The analysis of adjacent versus non-adjacent ratings in the eight columns was undertaken to determine if answers to neighboring food quality categories influenced one another. There was no significant difference between adjacent and non-adjacent ratings. This would be expected if the subject 'straight-lined'; i.e., gave all columns the same rating (from good to bad). This hypothesis also accounts for the high correlations among scales reported immediately above. To avoid this type of position-response bias on the NARADCOM card forms, the scales were designed so that a positive or negative answer was not physically in the same location on all four food attribute categories. To further encourage independence in the rating process, the categories were separated physically and enclosed in a 'box', vis-a-vis the continuous format used in the USATSA form.

NARADCOM Data

Four new rating card forms were developed for comparison with the USATSA form (See Appendix A, Figures 2, 3, 4, &5). Card #2 contains a single 7-point hedonic rating scale and a question asking if the subject selected the item. This extremely simple format was chosen to provide a basis for comparison with the more complex forms. If the complex forms did not provide additional information, then the simpler card would be preferable. The verbal anchors on the 7-point scale were carefully selected according to the guidelines given by Jones, Peryam and Thurstone (1955)[3], who showed that "good" and "bad"

[3]Jones, L. V., D. R. Peryam, and L. L. Thurstone. Food Research, 20(5): 512-520, (1955).

are well balanced around the mid-point 'neutral'. In addition, a space was left on the card titled 'comments'. It was felt that subjects might comment on technical aspects of the food which had contributed to their hedonic evaluation. A large number of comments would be viewed as indicating that the single response scale provided inadequate opportunity for the respondents to report their feelings.

Cards 3, 4, & 5 contain basically the same food quality evaluative questions (Temperature, Flavor, Portion Size, and Texture), the same selection question, and the same 5-point overall hedonic evaluation. Cards #3 and #4 differ only slightly. Card #3 has two additional questions which were taken from the USATSA form (Do you normally like this item? Do you like to eat here often? These questions were included for the purpose of comparing the data gathered with this new card and the USATSA card. Card #5 has seven verbal anchors in the food quality categories, rather than the five on cards 3 and 4. Statistically, a seven-point scale is preferable for data analysis because the scale allows greater separation of ratings. There was some question, however, as to whether the average subject could properly discriminate across the levels of a seven-point scale.

The choice of the category labels (temperature, etc.) was made after lengthy consultation with psychologists and food technologists at NARADCOM. The flavor and texture categories were intended to evaluate food quality aspects of the item, whereas the temperature

and portion size categories were directed at the food preparation aspects. For the purpose of this test, the verbal modifiers used for the flavor and texture categories were good/bad. It is clearly preferable to tailor these terms to the individual food being tested (such as tough/tender for meat) but none of these terms is appropriate for all the foods. The verbal modifiers used for temperature (hot/cold) and portion size (big/small) were chosen to convey more information than simply a good/bad statement. The five-point overall hedonic question which appeared on Cards 3, 4, & 5 was formulated using the guidelines mentioned earlier (Jones et al. 1955). Comment sections were available on Cards 3, 4, & 5 as they were on Cards 1 & 2.

Initial Testing of NARADCOM Cards at Fort Lee

The four rating card forms which were developed and tested at NARADCOM and the USATSA card were brought to Fort Lee for further testing. Two consolidated dining facilities were chosen for 5-day continuous testing. In one facility hamburgers were rated every day at lunch, and in the other, mashed potatoes were evaluated. No intentional adulterations were made to either food.

Although the numbers of both hamburger ratings and mashed potato ratings were small, it was reasonable to perform commonality analyses (a special case of step-wise multiple regression). The results of the analysis of mashed potato ratings and hamburger ratings were used as cross-validations. That is, conclusions were drawn about the performance of a rating card form if the analysis of that card

on both mashed potatoes and hamburger were in agreement. It was not possible to perform this analysis on Card Form 2 because it rated only one attribute.

The results of the commonality analyses and visual inspection favor neither Card Form 1 nor Card Form 2. These cards represent the extremes in terms of number of attributes to be rates (see Appendix A). Card 1 (the USATSA card) requires the subject to rate the food item on 8 attributes, some of which are abstract. There is reason to believe that many subjects find this large number of ratings to be confusing. In support of this, the data again showed generalized responses, that is, giving very similar responses to all questions without regard for the specific information requested - a "halo" effect. There were high correlations between ratings of attributes which are logically unrelated, such as that between Portion Size and Flavor or Texture and Serving Temperature.

Card Form 2, which requires the subject to merely give an overall hedonic rating of the food item is not satisfactory because many subjects made comments in areas of food quality which appear on the other four cards. This indicates the need to provide subjects with the opportunity to rate food items in terms of temperature, flavor, portion size, and texture, as well as in terms of their overall rating.

The results of the analysis of Card Forms 3, 4, and 5 were very similar. Thus fat, analyses indicated that all three were quite sensitive in measuring attitudes toward food items. Correlations

between such logically unrelated attributes as Flavor and Portion Size were low, and correlations between such logically related attributes as Flavor and Texture were moderately high (see Appendix B). Commonality analysis revealed that individual attribute ratings on Cards 3, 4, and 5 manifested more unique variation than individual attribute ratings on card 1.

A comparison of questions 2 and 3 on the USATSA card with the same questions on NARADCOM Card #3 revealed that the patterns of responding to these questions were essentially the same, i.e. the difference in format on Card #3 did not cause a change in response to questions 2 and 3.

Therefore, the analyses favored neither Rating Card 1 nor 2. The performance of Cards 3, 4 and 5 was approximately equal, but better than forms 1 & 2, in yielding lower correlations between logically unrelated categories. Following the comparative analysis of Cards 1 and 3, it was determined that Card 3 was longer needed for data collection.

Final Testing of NARADCOM Cards at Fort Lee

For the final evaluation of the rating cards, Card Form 3 was eliminated. It was nearly identical to Card Form 4, excepting the two additional behavior/attitude questions (earlier deemed superfluous). The use of only four card forms instead of five was expected to yield larger samples of people completing each card form, thus yielding more stable statistical results.

The results of commonality analyses were the same as the previous survey. The responses to the different food attributes as measured by Card 1 were highly correlated (see Appendix B). Attributes which were logically unrelated proved to be as highly correlated as attributes which were logically related. Very little unique variation was exhibited by any ratings and an enormous amount of all-rating overlap variation existed.

Cards 4 and 5, however, exhibited the desirable characteristics of a questionnaire. The food attribute ratings all had relatively large amounts of unique variation. There was some overlapping among attributes which were logically related and minimal overlap variation among logically unrelated attributes.

Because all cards were used to evaluate the same food, it was expected that the mean rating of an attribute as measured by one card form should be the same as the mean rating of that attribute as measured by the other card forms. It was also of interest to determine if ratings differed either by dining hall or by method of collection (picked up or handed in). Three-way ANOVAS were performed on each food attribute rating.

Neither card type, dining hall, nor method of card collection produced a significant difference on Temperature, Texture, Flavor, or Overall ratings. There was, however, a significant card by collection method interaction.[4] Follow-up analyses yielded inconclusive

---

[4]Analysis of Variance: $F(3, 140) = 3.18$. $P<.05$

results. While it was determined that people completing Card #5 and not bothering to hand it in were more sensitive to the small size of the cake, this finding was not considered practically important.

Because there was an attempt to make the Portion Size attribute of white cake inferior to the Temperature, Flavor, and Texture attributes, it was expected that if the cards were sensitive to the size reduction, an ANOVA performed on those four attributes with additional factors of dining hall, method of collection, and card type would yield a significant difference in attribute ratings. Although dining hall and method of collection showed no interaction, the card which had individual food attribute rating categories (1, 4 & 5) did show significant differences as expected.[5]

A planned comparison revealed that Portion Size ratings were significantly less favorable than Temperature, Flavor, and Texture ratings across all cards. This indicates that Cards 1, 4 and 5 are all sensitive to the small size of the cakes. This is a measure of validity for all of these cards.

It can be drawn from these analyses that method of card collection does not influence food ratings, nor does dining hall, as long as the food was the same in both dining halls, as the white cake was in this experiment.

In summary, the results of the white cake experiment tended to underline the conclusion drawn from the analysis of the hamburger

---

[5]Analysis of Variance: $F(3, 562) = 7.53$, $P<.05$.

and mashed potato data collected at Fort Lee in December 1975. Cards 1 and 2, the most extensive and the least extensive card forms, respectively, proved to be less desirable than Cards 4 and 5. Card 2 does not give enough information, and the attribute ratings of Card 1 do not demonstrate enough independence. Cards 4 and 5 again appear to be equally sensitive. Because they do appear to be equally useful, Card 5 is recommended. Card 5 allows the respondents to select from seven categories when rating each attribute as opposed to the five categories Card 4 allows. The use of Card 5 will allow the respondents to give more precise ratings of the foods.

CONCLUSIONS

1. Asking respondents on a food acceptance rating card to indicate whether they normally like the item is both unnecessary and redundant. All persons eating an item are entitled to an opinion, even though those normally liking the item generally give higher ratings.
2. The food acceptance rating card with eight rating scales yields too little independence of different scales. A hypothesis which may fit most of the data is that respondents are 'straight-lining', i.e., marking every column at the same level without property reading each scale separately. This yields high correlations between all pairs of scales whether logically related or not; therefore this card form is not recommended. A form which randomly reverses the evaluation polarity of the scales would inhibit 'straight-lining'.
3. A food acceptance rating card composed of a simple hedonic scale elicits a large number of written comments which would be measured by the scales on the other more complex card forms.

4. Both the card with eight scales and the cards with four scales show sensitivity when a food is intentionally adulterated on one response scale (small white cake study). This is interpreted to mean that these cards are valid measures of real differences in foods. The eight scale card is somewhat suspect, however, because adulteration of a food is evidenced in all eight categories to some degree, as was observed in the small white cake and roast beef studies.

5. The 7-point scales perform as well as the 5-point scales. Therefore, the 7-point scale is preferable because it will yield more information.

6. Therefore, because of (a) the lack of independence of the food rating categories on the eight category card, (b) the validity of Cards 1, 4, and 5, and (c) the statistical advantages of Card 5, (4 category, 7-point scale), Rating Card #5 was recommended for further use.

REFERENCES

1. Jacobs, H. L. and H. L. Meiselman. Customer morale and behavioral effectiveness: Accomplishments and goals of psychological studies of food service systems. In Proceedings, Third International Meeting, Foods for the Armed Forces. NARADCOM Technical Report 76-42-OTD. March, 1976.
2. Jones, L. V., D. R. Peryam, and L. L. Thurstone. Food Research, 20(5): 512-520, (1958).

APPENDIX A

Food Rating – Feedback Cards

Figure A–1 Original TSA/Fort Lee rating card containing 8 rating attributes

Figure A–2 NARADCOM Card #2; Simple 7-point hedonic scale

Figure A–3 NARADCOM Card #3; 5-point scale, four category card with behavior/attitude questions

Figure A–4 NARADCOM Card #4; 5-point scale, four category card

Figure A–5 NARADCOM Card #5; 7-point, four category card

INSTRUCTIONS: This card will let you comment on an item on today's menu. If you didn't select the item, check the NO box and turn in the card as you leave. If you did select the item, check the YES box. Then score the item you selected AFTER you eat it by checking the box best describing your feeling.

DID YOU SELECT *WHITE CAKE*? ☐ YES ☐ NO | DO YOU NORMALLY LIKE THIS ITEM? ☐ YES ☐ NO | DO YOU LIKE TO EAT HERE OFTEN? ☐ YES ☐ NO | CARD C

RATE FOR →	APPEARANCE ON YOUR PLATE	SERVING TEMPERATURE	TEXTURE	FLAVOR	MOISTNESS	COOK'S PREPARATION OF ITEM	PORTION SIZE	GOES WITH THE REST OF YOUR MEAL
GOOD	☐	☐	☐	☐	☐	☐	☐	☐
SOMEWHAT GOOD	☐	☐	☐	☐	☐	☐	☐	☐
MEDIUM	☐	☐	☐	☐	☐	☐	☐	☐
SOMEWHAT BAD	☐	☐	☐	☐	☐	☐	☐	☐
BAD	☐	☐	☐	☐	☐	☐	☐	☐

TSA Form 6086 6 Jun 14

PLEASE COMMENT HERE:

FIGURE A-1 ORIGINAL TSA/FORT LEE RATING CARD CONTAINING 8 RATING ATTRIBUTES

DID YOU SELECT *WHITE CAKE*? ☐ YES (continue answering) ☐ NO (STOP & return card) C

After you have tasted the *WHITE CAKE*, please rate it by marking a box on the scale below.

Very Good	Good	Slightly Good	Neutral	Slightly Bad	Bad	Very Bad
☐	☐	☐	☐	☐	☐	☐

TSA Form 678 30 Oct 15

COMMENTS:

FIGURE A-2 NARADCOM CARD #2 SIMPLE 7-POINT HEDONIC SCALE

DID YOU SELECT *WHITE CAKE*? ☐ YES (continue answering) ☐ NO (STOP & return card) C

DO YOU NORMALLY LIKE THIS ITEM? ☐ YES ☐ NO DO YOU LIKE TO EAT HERE OFTEN? ☐ YES ☐ NO

After you have tasted the *WHITE CAKE*, please rate it for each of the following characteristics by checking one box in each category.

Temperature	Flavor	Portion Size	Texture
Too Hot ☐	Good Flavor ☐	Too Big ☐	Bad Texture ☐
Slightly Too Hot ☐	Slightly Good Flavor ☐	Slightly Too Big ☐	Slightly Bad Texture ☐
Just Right ☐	Neutral Flavor ☐	Just Right ☐	Neutral Texture ☐
Slightly Too Cold ☐	Slightly Bad Flavor ☐	Slightly Too Small ☐	Slightly Good Texture ☐
Too Cold ☐	Bad Flavor ☐	Too Small ☐	Good Texture ☐

TSA Form 936 30 Oct 15

Considering everything, how was the *WHITE CAKE*?

Good ☐ Slightly Good ☐ Neutral ☐ Slightly Bad ☐ Bad ☐

COMMENTS:

FIGURE A-3 NARADCOM CARD #3; 5-POINT SCALE, FOUR CATEGORY CARD WITH BEHAVIORAL/ATTITUDE QUESTIONS.

DID YOU SELECT *WHITE CAKE* ? ☐ YES (continue answering) ☐ NO (STOP & return card) **C**

After you have tested the *WHITE CAKE*, please rate it for each of the following characteristics by checking one box in each category.

Temperature	Flavor	Portion Size	Texture
Too Hot ☐	Good Flavor ☐	Too Big ☐	Bad Texture ☐
Slightly Too Hot ☐	Slightly Good Flavor ☐	Slightly Too Big ☐	Slightly Bad Texture ☐
Just Right ☐	Neutral Flavor ☐	Just Right ☐	Neutral Texture ☐
Slightly Too Cold ☐	Slightly Bad Flavor ☐	Slightly Too Small ☐	Slightly Good Texture ☐
Too Cold ☐	Bad Flavor ☐	Too Small ☐	Good Texture ☐

TEA Form 935, 30 Oct 75

Considering everything, how was the *WHITE CAKE*?

Good ☐ Slightly Good ☐ Neutral ☐ Slightly Bad ☐ Bad ☐

COMMENTS:

FIGURE A-4 NARADCOM CARD #4 5-POINT SCALE, FOUR CATEGORY CARD

DID YOU SELECT *WHITE CAKE* ? ☐ YES (continue answering) ☐ NO (STOP & return card) **C**

After you have tested the *WHITE CAKE*, please rate it for each of the following characteristics by checking one box in each category.

Temperature	Flavor	Portion Size	Texture
Much Too Hot ☐	Very Good Flavor ☐	Much Too Big ☐	Very Bad Texture ☐
Too Hot ☐	Good Flavor ☐	Too Big ☐	Bad Texture ☐
Slightly Too Hot ☐	Slightly Good Flavor ☐	Slightly Too Big ☐	Slightly Bad Texture ☐
Just Right ☐	Neutral Flavor ☐	Just Right ☐	Neutral Texture ☐
Slightly Too Cold ☐	Slightly Bad Flavor ☐	Slightly Too Small ☐	Slightly Good Texture ☐
Too Cold ☐	Bad Flavor ☐	Too Small ☐	Good Texture ☐
Much Too Cold ☐	Very Bad Flavor ☐	Much Too Small ☐	Very Good Texture ☐

TEA Form 936, 30 Oct 75

Considering everything, how was the *WHITE CAKE*?

Good ☐ Slightly Good ☐ Neutral ☐ Slightly Bad ☐ Bad ☐

COMMENTS:

FIGURE A-5 NARADCOM CARD #5 7-POINT SCALE, FOUR CATEGORY CARD

APPENDIX B

Descriptive Statistics on Food Ratings

I. Means and Standard Deviations

B–1 – Mashed Potatoes
B–2 – Hamburger Patties
B–3 – Small White Cake

II. Rating Intercorrelation

B–4 – Correlation Matrices for Rating Card 1; Mashed Potatoes and Hamburger Data

B–5 – Correlation Matrices for Rating Card 3; Mashed Potatoes and Hamburger Data

B–6 – Correlation Matrices for Rating Card 4; Mashed Potatoes and Hamburger Data

B–7 – Correlation Matrices for Rating Card 5; Mashed Potatoes and Hamburger Data

B–8 – Correlation Matrix for Rating Card 1; Small White Cake Data

B–9 – Correlation Matrices for Rating Cards 4 and 5; Small White Cake Data

TABLE B–1

DESCRIPTIVE STATISTICS ON RATINGS GIVEN TO MASHED POTATOES

Attributes	Card 1 Mean	Card 1 Std Dev	Card 1 N	Card 2 Mean	Card 2 Std Dev	Card 2 N	Card 3 Mean	Card 3 Std Dev	Card 3 N	Card 4 Mean	Card 4 Std Dev	Card 4 N	Card 5 Mean	Card 5 Std Dev	Card 5 N
Normally like	1.05	0.23	19				1.18	0.39	17						
Like to eat here	1.06	0.23	20				1.23	0.44	13						
Appearance	4.55	0.96	24												
Temperature	4.27	0.98	25				4.45	1.10	22	3.93	1.36	30	5.94*	1.44	32
Texture	4.09	1.15	24				3.70	1.26	23	3.57	1.43	30	4.33*	1.76	33
Flavor	4.14	1.12	23				3.87	1.32	23	3.48	1.50	31	4.82*	1.69	33
Moisture	3.68	1.24	24												
Cook's Preparation	4.14	1.08	25												
Portion size	4.14	1.08	25				4.30	1.29	23	2.93	1.73	29	5.81*	2.26	33
Goes with meal	4.50	0.96	24												
Overall				4.70*	2.25	20	4.22	1.24	23	3.71	1.49	31	3.72	1.37	32

*Responses to this question range from 1 (very bad) to 4 (neutral) to 7 (very good).
Responses to all other questions range from 1 (very bad) to 3 (neutral) to 5 (very good).

TABLE B–2

DESCRIPTIVE STATISTICS ON RATINGS GIVEN TO HAMBURGER PATTIES

Attirbutes	Card 1 Mean	Card 1 Std Dev	Card 1 N	Card 2 Mean	Card 2 Std Dev	Card 2 N	Card 3 Mean	Card 3 Std Dev	Card 3 N	Card 4 Mean	Card 4 Std Dev	Card 4 N	Card 5 Mean	Card 5 Std Dev	Card 5 N
Normally like	1.05	0.05	19				1.11	0.32	18						
Like to eat here	1.06	.06	18				1.17	0.38	18						
Appearance	4.55	0.93	22												
Temperature	4.27	0.97	22				4.41	1.18	17	4.19	1.38	27	6.33*	1.27	24
Texture	4.09	1.32	22				4.00	1.00	17	3.81	1.47	27	5.04*	1.55	23
Flavor	4.13	1.27	22				3.71	1.16	17	4.00	1.33	27	5.64*	1.22	22
Moisture	3.68	1.56	22												
Cook's preparation	4.14	1.17	22												
Portion size	4.14	1.17	22				3.11	1.80	17	4.19	1.49	27	5.75*	1.75	24
Goes with meal	4.50	0.93	22												
Overall				5.50*	1.76	18	4.24	1.5	17	3.92	1.38	24	4.27	1.08	22

*Responses to this question range from 1 (very bad) to 4 (neutral) to 7 (very good)
Responses to all other questions range from 1 (very bad) to 3 (neutral) to 5 (very good).

1037

1141

1142

TABLE B–3

DESCRIPTIVE STATISTICS ON RATINGS GIVEN TO
SMALL WHITE CAKE – COLLECTED FEB 1976 AT FORT LEE

Attributes	Card 1 Mean	Card 1 Std Dev	Card 1 N	Card 2 Mean	Card 2 Std Dev	Card 2 N	Card 4 Mean	Card 4 Std Dev	Card 4 N	Card 5 Mean	Card 5 Std Dev	Card 5 N
Normally like	1.33	0.47	40									
Like to eat here	1.33	0.47	40									
Appearance	4.04	1.21	49									
Temperature	3.93	1.30	44				4.04	1.32	56	5.37*	2.28	43
Texture	3.69	1.47	45				3.32	1.41	59	4.41*	1.82	44
Flavor	3.82	1.51	45				3.69	1.34	59	5.15*	1.69	48
Moisture	3.57	1.52	44									
Cook's preparation	3.93	1.44	43									
Portion size	3.07	1.68	44				2.62	1.61	58	3.09*	2.04	47
Goes with meal	3.95	1.48	42									
Overall				5.12*	1.94	49	3.86	1.29	56	3.84	1.27	43

*Responses to this question range from 1 (very bad) to 4 (neutral) to 7 (very good).
Responses to all other questions range from 1 (very bad) to 3 (neutral) to 5 (very good).

1038

1143

1144

TABLE B–4

CORRELATION MATRICES FOR RATING CARD 1;
MASHED POTATOES AND HAMBURGER DATA

Mashed Potatoes – Card 1

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Like	1.0000	.1359	.1450	.0171	.0000	.0000	–.1423	–.5533	–.0933	.1685
2 Eat here	.1359	1.0000	–.6262	–.5967	–.5626	–.5099	–.7553	–.4994	–.4122	–.6029
3 Appearance	.1450	–.6262	1.0000	.7955	.8851	.8382	.6450	.4617	.5107	.7514
4 Temperature	.0171	–.5967	.7955	1.0000	.7133	.8249	.8011	.6059	.5273	.7804
5 Texture	.0000	–.5626	.8851	.7133	1.0000	.8643	.7795	.6396	.5654	.7971
6 Flavor	.0000	–.5099	.8382	.8249	.8643	1.0000	.7017	.7625	.5745	.7668
7 Moisture	–.1423	–.7553	.6450	.8011	.7795	.7017	1.0000	.6482	.5708	.7033
8 Cook's preparation	–.5533	–.4994	.4617	.8059	.6396	.7625	.6482	1.0000	.5589	.6032
9 Portion size	–.0933	–.4122	.5107	.5273	.5654	.5745	.5708	.5589	1.0000	.5950
10 Goes with meal	.1685	–.6029	.7514	.7804	.7971	.7668	.7033	.6032	.5950	1.0000

1039

Hamburger – Card 1

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Like	1.0000	.0000	–.7003	–.1472	–.1074	–.3368	–.1617	–.3508	.2047	.1313
2 Eat here	.0000	1.0000	.1296	–.5115	.1795	–.1059	–.1732	–.3961	–.3473	–.7001
3 Appearance	–.7003	.1296	1.0000	.6897	.5979	.6315	.4284	.6567	.5195	.6161
4 Temperature	–.1472	–.5115	.6897	1.0000	.6073	.7384	.4997	.6784	.7678	.8029
5 Texture	–.1074	.1785	.5879	.6073	1.0000	.8356	.8503	.7160	.7160	.6011
6 Flavor	–.3368	–.1059	.6315	.7384	.8356	1.0000	.8453	.8835	.7661	.6807
7 Moisture	–.1617	–.1732	.4204	.4997	.8503	.8453	1.0000	.8790	.7029	.5736
8 Cook's preparation	–.3508	–.3961	.6567	.6784	.7160	.8835	.8790	1.0000	.7967	.7535
9 Portion size	.2047	–.3473	.5195	.7678	.7160	.7881	.7029	.7967	1.0000	.8448
10 Goes with meal	.1313	–.7001	.6161	.8029	.6011	.6807	.5736	.7535	.8448	1.0000

TABLE B-5

CORRELATION MATRICES FOR RATING CARD 3;
MASHED POTATOES AND HAMBURGER DATA

Mashed Potatoes – Card 3

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Normally like	1.0000	−.2000	.2744	.0977	−.0318	.2486	.1330			
2 Eat here	−.2000	1.0000	.3000	−.2865	−.6260	−.3156	−.2287			
3 Temperature	.2774	.3000	1.0000	.3945	.3705	.4980	.5321			
4 Flavor	.0977	−.2865	.3945	1.0000	.5278	.6837	.8155			
5 Size	−.0318	−.6260	.3705	.5278	1.0000	.5335	.5936			
6 Texture	.2488	−.3156	.4980	.6837	.5335	1.0000	.8287			
7 Overall rating	.1330	−.2287	.5321	.8155	.5836	.8287	1.0000			

Hamburger Patty – Card 3

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Normally like	1.0000	.7906	.1289	−.6011	−.3034	−.5154	−.7333			
2 Eat here	.7906	1.0000	.1883	−.5535	−.2339	−.1882	−.5138			
3 Temperature	.1289	.1883	1.0000	.2318	.1530	.4253	.5320			
4 Flavor	−.6011	−.5535	.2318	1.0000	.0775	.7543	.6057			
5 Size	−.3034	−.2339	.1530	.0775	1.0000	.2085	.2747			
6 Texture	−.5154	−.1882	.4253	.7543	.2085	1.0000	.6344			
7 Overall rating	−.7333	−.5138	.5320	.6057	.2747	.6344	1.0000			

TABLE B–6

CORRELATION MATRICES FOR RATING CARD 4; MASHED POTATOES AND HAMBURGER DATA

Mashed Potatoes – Card 4

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Temperature	1.0000	.4895	.3937	.1830	.3945					
2 Flavor	.4895	1.0000	.0928	.7504	.8472					
3 Size	.3937	.0928	1.0000	–.0402	.0364					
4 Texture	.1830	.7504	–.0402	1.0000	.7576					
5 Overall rating	.3945	.8472	.0364	.7576	1.0000					

1041

Hamburger Patty – Card 4

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Temperature	1.0000	.5834	.5577	.3006	.7698					
2 Flavor	.5834	1.0000	.8126	.5907	.8108					
3 Size	.5577	.8126	1.0000	.6296	.7848					
4 Texture	.3006	.5907	.6296	1.0000	.4870					
5 Overall rating	.7698	.8108	.7848	.4870	1.0000					

1149

1150

TABLE B–7

CORRELATION MATRICES FOR RATING CARD 5;
MASHED POTATOES AND HAMBURGER DATA

Mashed Potatoes – Card 5

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Temperature	1.0000	.2313	.1105	.0918	.1414					
2 Flavor	.2313	1.0000	.3541	.6945	.6082					
3 Size	.1105	.3541	1.0000	.3931	.4047					
4 Texture	.0918	.6945	.3931	1.0000	.6968					
5 Overall rating	.1414	.6082	.4047	.6968	1.0000					

Hamburger Patty – Card 5

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Temperature	1.0000	.4578	.2339	.3326	.6170					
2 Flavor	.4578	1.0000	.1659	.1173	.3125					
3 Size	.2339	.1659	1.0000	−.0115	.1529					
4 Texture	.3326	.1173	−.0115	1.0000	.8163					
5 Overall rating	.6170	.3125	.1529	.8163	1.0000					

TABLE 8

CORRELATION MATRIX FOR RATING CARD 1:
SMALL WHITE CAKE DATA

Small White Cake — Card 1

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Dining hall	1.0000	.4215	−.1234	.0912	.0308	.0839	−.0375	−.1112	.0803	−.1755
2 Collected/handed in	.4215	1.0000	.2213	.0117	.0885	.0742	.0862	.1762	.0666	.0691
3 Normally like item	−.1234	−.2213	1.0000	.2357	−3282	−.3318	−.3930	−.5337	−.5356	−.4950
4 Usually eat here	.0912	.0117	.2357	1.0000	−.4999	−.4010	−.4112	−.4631	−.5356	−.4547
5 Appearance	.0308	.0885	−.3282	−.4999	1.0000	.7343	.6372	.7303	.7792	.7555
6 Temperature	−.0839	.0742	−.3318	−.4010	.7343	1.0000	.7560	.6980	.6804	.7194
7 Texture	−.0375	.0862	−.3930	−.4112	.6372	.7560	1.0000	.7245	.8435	.7311
8 Flavor	−.1112	.1762	−.5357	−.4631	.7303	.6980	.7245	1.0000	.8554	.9024
9 Moisture	−.0803	.0666	−.5356	−.5356	.7792	.6804	.8435	.8554	1.0000	.7816
10 Cooks preparation	−.1755	.0691	−.4950	−.4547	.7555	.7194	.7311	.9024	.7816	1.0000
11 Size	.1478	.0897	−.4676	−.3794	.6546	.5567	.6884	.6962	.7350	.6160
12 Goes with meal	.0667	.1274	−.2517	−.3597	.7496	.7430	.8397	.7068	.7323	.7303

1043

Small White Cake — Card 1

Correlation Matrix

Variable	11	12
1 Dining hall	.1478	.0667
2 Collected/handed in	.0897	.1274
3 Normally like item	−.4676	−.3517
4 Usually eat here	−.3794	−.3597
5 Appearance	.6546	.7496
6 Temperature	.5567	.7430
7 Texture	.6884	.8397
8 Flavor	.6962	.7068
9 Moisture	.7350	.7323
10 Cooks preparation	.6160	.7303
11 Size	1.0000	.7649
12 Goes with meal	.7649	1.0000

TABLE B–9

CORRELATION MATRICES FOR RATING CARDS 4 and 5;
SMALL WHITE CAKE DATA

Small White Cake – Card 4

Variable	Correlation Matrix 1	2	3	4	5	6	7
1 Dining hall	1.0000	.3640	.0029	−.0274	−.0986	−.0580	.0405
2 Collected-hand in	.3640	1.0000	−.0127	−.1519	.0513	−.3600	.0523
3 Temperature	.0029	−.0127	1.0000	.4724	.3793	.4165	.8686
4 Flavor	−.0274	−.1519	.4724	1.0000	.2281	.5640	.7628
5 Size	−.0986	.0513	.3793	.2281	1.0000	.2444	.3980
6 Texture	−.0580	−.3600	.4165	.5640	.2444	1.0000	.6149
7 Overall	.0405	.0523	.6686	.7628	.3980	.6149	1.0000

Small White Cake – Card 5

Variable	Correlation Matrix 1	2	3	4	5	6	7
1 Dining hall	1.0000	.1839	.0390	−.0390	.0416	.0380	.2780
2 Collected-handed in	.1839	1.0000	−.1110	−.0236	−.1716	−.0822	−.0586
3 Temperature	.0390	−.1110	1.0000	.3291	.1911	.3710	.3041
4 Flavor	−.0390	−.0236	.3291	1.0000	.3391	.6484	.6942
5 Size	.0416	−.1716	.1911	.3391	1.0000	.4342	.4639
6 Texture	.0380	−.0822	.3710	.6484	.4342	1.0000	.7760
7 Overall	.2780	−.0586	.3041	.6942	.4639	.7760	1.0000

TABLE B–8

CORRELATION MATRIX FOR RATING CARD 1; SMALL WHITE CAKE DATA

Small White Cake – Card 1

Correlation Matrix

Variable	1	2	3	4	5	6	7	8	9	10
1 Dining hall	1.0000	.4215	–.1234	.0912	.0308	.0839	–.0375	–.1112	.0803	–.1755
2 Collected/handed in	.4215	1.0000	.2213	.0117	.0885	.0742	.0862	.1762	.0666	.0691
3 Normally like item	–.1234	–.2213	1.0000	.2357	–3282	–.3318	–.3930	–.5337	–.5356	–.4950
4 Usually eat here	.0912	.0117	.2357	1.0000	–.4999	–.4010	–.4112	–.4631	–.5356	–.4547
5 Appearance	.0308	.0885	–.3282	–.4999	1.0000	.7343	.6372	.7303	.7792	.7555
6 Temperature	–.0839	.0742	–.3318	–.4010	.7343	1.0000	.7560	.6980	.6804	.7194
7 Texture	–.0375	.0862	–.3930	–.4112	.6372	.7560	1.0000	.7245	.8435	.7311
8 Flavor	–.1112	.1762	–.5357	–.4631	.7303	.6980	.7245	1.0000	.8554	.9024
9 Moisture	–.0803	.0666	–.5356	–.5356	.7792	.6804	.8435	.8554	1.0000	.7816
10 Cooks preparation	–.1755	.0691	–.4950	–.4547	.7555	.7194	.7311	.9024	.7816	1.0000
11 Size	.1478	.0897	–.4676	–.3794	.6546	.5567	.6884	.6962	.7350	.6160
12 Goes with meal	.0667	.1274	–.2517	–.3597	.7496	.7430	.8397	.7068	.7323	.7303

Small White Cake – Card 1

Correlation Matrix

Variable	11	12
1 Dining hall	.1478	.0667
2 Collected/handed in	.0897	.1274
3 Normally like item	–.4676	–.3517
4 Usually eat here	–.3794	–.3597
5 Appearance	.6546	.7496
6 Temperature	.5567	.7430
7 Texture	.6884	.8397
8 Flavor	.6962	.7068
9 Moisture	.7350	.7323
10 Cooks preparation	.6160	.7303
11 Size	1.0000	.7649
12 Goes with meal	.7649	1.0000

1156

1157

TABLE B–9

CORRELATION MATRICES FOR RATING CARDS 4 and 5;
SMALL WHITE CAKE DATA

Small White Cake – Card 4

	Correlation Matrix						
Variable	1	2	3	4	5	6	7
1 Dining hall	1.0000	.3640	.0029	–.0274	–.0986	–.0580	.0405
2 Collected-hand in	.3540	1.0000	–.0127	–.1519	.0513	–.3600	.0523
3 Temperature	.0029	–.0127	1.0000	.4724	.3793	.4165	.8686
4 Flavor	–.0274	–.1519	.4724	1.0000	.2281	.5640	.7628
5 Size	–.0986	.0513	.3793	.2281	1.0000	.2444	.3980
6 Texture	–.0580	–.3600	.4165	.5640	.2444	1.0000	.6149
7 Overall	.0405	.0523	.6686	.7628	.3980	.6149	1.0000

Small White Cake – Card 5

	Correlation Matrix						
Variable	1	2	3	4	5	6	7
1 Dining hall	1.0000	.1839	.0390	–.0390	.0416	.0380	.2780
2 Collected-handed in	.1839	1.0000	–.1110	–.0236	–.1716	–.0822	–.0586
3 Temperature	.0390	–.1110	1.0000	.3291	.1911	.3710	.3041
4 Flavor	–.0390	–.0236	.3291	1.0000	.3391	.6484	.6942
5 Size	.0416	–.1716	.1911	.3391	1.0000	.4342	.4639
6 Texture	.0380	–.0822	.3710	.6484	.4342	1.0000	.7760
7 Overall	.2780	–.0586	.3041	.6942	.4639	.7760	1.0000

1158

1159

Worker Job Satisfaction and Opinion in Military Food Service System Design

Lawrence E. Symington
Behavioral Sciences Division
Food Sciences Laboratory
U.S. Army Natick Research and Development Command

Presented to: Military Testing Association
October, 1976

Worker Job Satisfaction and Opinion in Military Food Service System Design

The input of behavioral science into the analysis of food service systems in the military has been traditionally oriented toward customer food preferences and customer opinions of the food service system. However, a trained, motivated food service worker is as important to the success of a food service system as the quality of the food, the equipment, and a facility which is well designed. In recent Department of Defense Food programs oriented toward systems analysis and redesign of various food service systems, the workers' opinions about systems changes and the resultant levels of job satisfaction have played a more important role in the overall assessment of such changes.

Assessment of any system change requires the collecting of customer and worker data both before and after shifting systems, in a setting where other changes are minimized. In the instance of the food service systems which have been analyzed, this has involved the surveying/interviewing of workers prior to the implementation of system change in order to assess their opinions of the existing system and the impending changes. In addition, their satisfaction with their job was assessed using a traditional industry job satisfaction instrument, the Job Description Index (Smith et al., 1969). These same procedures were followed at different time intervals after systems changes took place.

Job Satisfaction of the Food Service Worker in the Traditional Air Force Food Service System. The initial effort of the Behavioral Sciences Division, Food Sciences Laboratory, U.S. Army Natick Research and Development Command in the worker area was in an Air Force study to evaluate and redesign

the garrison food service system (Branch, et al., 1974,. Symington and Meiselman, 1975b). Travis Air Force Base, California was selected as the principal study site, with Minot Air Force Base, North Dakota and Homestead Air Force Base, Florida being chosen as ancillary sites. In this effort surveys, interviews and the Job Description Index (JDI) were administered to 157 military and 120 civil service food service workers before system changes were made.

The same procedures were instituted following the system changes made at the primary site, Travis AFB; and the results reported (Symington and Meiselman, 1975a). However, the focus of the present discussion is on the JDI data from the pre-change sample of workers, data which indicate where job satisfaction problems exist for the worker in the traditional military food service system.

The JDI is a standard paper and pencil instrument which measures satisfaction within five areas of a job: the work itself, the supervision, the co-workers on the job, the opportunities for promotion, and the pay. Before discussing the responses of the workers to this instrument, a brief explanation of the scoring should be undertaken.

Each of the five areas of the JDI is evaluated by responses to a list of words or descriptive phases (18 words and phrases each for work, supervision, and co-workers; nine each for pay and promotion). Figure 1 shows the format and four of the words/phrases from the work scale. The respondent circles the "Y" (yes) or "N" (no) to indicate whether or not the word/phrase describes his job. He circles "?" for those items which he does not understand or on which he cannot decide.

Based on a large number of respondents who were asked to describe the best and worst possible jobs for themselves, the developers of the JDI

Figure 1. Format for the Work Scale
of the Job Description Index (JDI)

Work

Fascinating	Y	N	?
Routine	Y	N	?
Boring	Y	N	?
Good	Y	N	?

(Smith et al., 1969) determined which response should be scored as "satisfied" for each item. For example, in Figure 1 "routine" and "boring" are scored in the satisfied direction if the individual responds, "N"; and "fascinating" and "good" are scored in the satisfied direction if he answers, "Y".

Traditional scoring methods would probably suggest a scoring of 2 for a "satisfied" answer, 1 for a "?" answer, and 0 for a "dissatisfied" answer. Smith, however, has scored "satisfied" answers as 3, "dissatisfied" answers as 0, and "?" asnwers as 1. This departure from traditional methodology was based on the response of the sample mentioned above where it was concluded that the "?" response was more indicative of dissatisfaction than of satisfaction. For each scale or area of the JDI, the range of possible scores is from 0 to 54, with scores on each word or phrase being summed for the work, supervision, and co-worker scales; and summed, then doubled for the pay and promotion scales.

Table 1 shows the overall mean responses of the military food service workers to the five scales of the JDI for the three Air Force bases, and also provides norms from a large, non-food service civilian male sample (Smith, et al., 1969). This civilian normative sample was drawn from private business organizations of fifty or more employees and from all levels within these organizations. The companies included production plants, retail stores, and research organizations, among others.

Overall, it can be seen that the workers as a whole were least satisfied with promotions, pay, and the work itself; and more satisfied with the people with whom they work and the supervision. (Mean responses to the five scales were significantly different from one another, see Table 2). On the other hand, ratings of promotion, the area with which the Air Force workers

Table 1. Mean Responses to the JDI at the Three Air Force Bases*

Scale	Travis AFB		Homestead AFB		Minot AFB		Mean for the Three AFB's		Norms**
	Military	Civilian	Military	Civilian	Military	Civilian	Military	Civilian	
Supervision	37.30	39.06	40.67	38.96	38.72	41.55	38.90	39.00	41.10
Co-workers	37.96	35.12	33.83	39.87	33.16	36.29	34.98	37.09	43.49
Work	26.07	28.73	21.83	32.00	22.25	29.10	23.38 b	29.94	36.57
Promotion	22.96	21.55	26.57	17.08	27.53	17.68	25.69 b	18.77	22.06
Pay[a]	20.89	29.27	16.91	23.00	25.97	28.97	21.26 b	27.08	29.90
Average	29.04	30.75	27.96	30.18	29.53	30.72	28.84	30.55	34.62

a. 3 bases significantly different on this scale (military and civilian combined)

b. Civilians significantly different from military on this scale (3 bases combined)

* Scores on scales of JDI range from 0 (lowest satisfaction) to 54 (highest satisfaction)

** Smith, et al., 1969

1050

1165

1166

Table 2. Statistical Differences Among All Five JDI Scales - Data Combined from All Three Bases

Work - Supervision	$X^2 = 181.11$, 16 df, $p < .001$
Work - People	$X^2 = 97.33$, 16 df, $p < .001$
Work - Pay	$X^2 = 61.16$, 16 df, $p < .001$
Work - Promotion	$X^2 = 137.08$, 17 df, $p < .001$
Supervision - People	$X^2 = 34.88$, 15 df, $p < .01$
Supervision - Pay	$X^2 = 138.19$, 17 df, $p < .001$
Supervision - Promotion	$X^2 = 161.56$, 17 df, $p < .001$
People - Pay	$X^2 = 83.16$, 17 df, $p < .001$
People - Promotion	$X^2 = 108.88$, 17 df, $p < .001$
Pay - Promotion	$X^2 = 56.24$, 18 df, $p < .001$

were least satisfied, approach, and in some instances, exceed the civilian norm. Conversely, within the Air Force data, satisfaction with work fell in the middle of the five scales, even though it is well below the comparable normative score. Satisfaction with pay, supervision, and co-workers at these bases is also less than the norm, although not to as great an extent. Since military pay is not within the purview of food service, these data suggest that anyone interested in improving satisfaction of the food service workers should concentrate on two areas, the work itself and co-worker relations, with a tertiary effort in the area of promotion.

The same table provides data which allow an examination of possible differences in satisfaction across the bases surveyed. When the data from all five satisfaction scales are combined, no significant differences appear; and when the data are broken down by scale, ratings on only one scale, pay, are significantly different across bases, with workers at Minot being most satisfied with their pay and workers at Homestead least satisfied ($X^2 = 38.00$, 16 df, $p < .01$). It is interesting to speculate on whether or not the relative isolation or cost of living of Minot might be the factor underlying this relative satisfaction. In any case, such differences among bases demonstrate the risk in using a single installation to evaluate any food service system.

When all five scales are combined there is no significant difference in satisfaction between civilian and military food service workers. Significant differences do appear, however, when the data are analyzed by scale over all three bases. The civilian workers were more satisfied with the work itself ($X^2 = 26.62$, 8 df, $p < .001$) and with their pay ($X^2 = 15.53$, 8 df, $p < .05$) than were their military co-workers. On the other hand, the military workers were more satisfied with promotion ($X^2 = 9.55$, 2 df, $p < .01$). The relative dissatisfaction on the part of the military workers with their work merits

serious consideration in terms of potential retention problems. The lower satisfaction with pay for military personnel might be partially a function of working closely with more highly paid civilians, although additional data bearing on satisfaction with pay for airmen of similar ranks would be necessary to verify such speculation.

Use of the Job Description Index (JDI) and Interview Data to Evaluate the BAS/A La Carte Food System. While the data discussed above or the basis for recommendations to the Air Force about the general state of food service worker satisfaction, the same types of data have been used in evaluating experimental food systems. One such experimental system currently being tested or planned for test by all four services is BAS/A La Carte (or, as the Navy labels it, CASH/A La Carte). The traditional food service system for all four services provides some of the enlisted personnel a meal card which they show to obtain their meals (rations in kind, RIK). The other enlisted personnel, mostly married or senior level personnel, receive a monthly subsistence allowance, Basic Allowance for Subsistence (BAS), or Commuted Rations, (COMRATS). When an individual on BAS (or COMRATS) eats in the dining facility, he pays a set amount for a full meal regardless of what he actually selects.

In the BAS/A La Carte tests, all enlisted personnel on the installation receive the food allowance (BAS or COMRATS)rather than the authorization to eat in the dining hall at no cost. In addition, all pay cash for food in the dining hall on an item-by-item rather than on a whole-meal basis. Hence, the names BAS/A LaCarte or CASH/A La Carte. While the traditional food service system offers limited food choice (two entrees) the new system greatly expands the variety of entrees as well as other components of the meal.

To date both the Air Force and Navy have tested this system (at Loring

Air Force Base, Maine and Alameda Naval Air Station, California). In both tests the general response of the customers has been similar, being extremely positive about receiving BAS or COMRATS, and also positive about item pricing and the variety offered. Results from the JDI and interviews with the military food service workers, however, have indicated marked differences in the opinion of Air Force and Navy cooks about the new system.

Table 3 shows the mean response of Pre- and Post-Test (before and after implementation of BAS or CASH/A La Carte) Loring AFB and NAS Alameda workers to the work scale of the JDI. (At both bases the Pre-Test and Post-Test samples were essentially identical, but differed slightly due to leave, transfers, etc.) At Loring the Post-Test mean was higher than the Pre-Test mean, although the difference was not statistically significant, suggesting that the Air Force cooks found the work in the new system at least no worse than their work in the old system. However, the Navy Mess Management Specialists at NAS Alameda responded differently, with the Post-Test workers having a statistically significant lower mean than the Pre-Test workers ($X^2 = 6.39$, 2 df, $p < .05$). NAS Alameda workers, then, were significantly less satisfied with the CASH/A La Carte system as implemented at NAS Alameda than they were with the traditional system.

Responses to the interviews used at both installations verified these data and also suggested reasons why Loring workers were satisfied with the new system, and Alameda workers, dissatisfied. One of the interview questions asked the workers to express their preference for the BAS (CASH)/A La Carte or traditional system. The responses to this question supported the JDI responses with 83% of the Loring workers preferring the new system to some degree and 56% of the Alameda workers preferring the traditional

Table 3. Mean JDI Work Scale Scores* for Pre and Post-Test Loring AFB and NAS Alameda Military Food Service Workers

Loring AFB		NAS Alameda	
Pre (N=25)	Post (N=30)	Pre (N=30)	Post (N=25)
24.84	27.53	29.20 **	21.54

* 0 = lowest satisfaction, 54 = highest satisfaction

** $X^2 = 6.39$, 2 df, $p < .05$

system (Table 4). The difference between the two bases is statistically significant ($X^2 = 16.95$, 2 df, $p < .001$). At Loring AFB the 63% who "extremely preferred" BAS/A La Carte were fairly evenly distributed among "workers" and "supervisors". At NAS Alameda, however, five of the six who "extremely preferred" CASH/A La Carte were "supervisors" (food service officer, leading MS, galley captain, or watch captain). So, at Alameda the dissatisfied personnel tended to be at the working rather than supervision level.

Additional open ended interview questions concerning the good and bad aspects of the new system pointed up some important differences between Loring and Alameda workers. Loring AFB workers felt that their jobs were harder under BAS/A La Carte, but maintained that the greater variety offered increased their opportunities to cook. In the traditional system, they claimed, civilian wage grade cooks had done the actual food preparation while military workers had been mostly restricted to cleaning and serving duties. In the BAS/A LaCarte system they were preparing food which, in most instances, was their image of what a cook should do. This was not the case for the NAS Alameda workers who had also cooked in the traditional system since there were no civilian cooks. Their main perception of the new system was oriented toward longer hours and heavier workloads. The change had not "made them cooks" as it reportedly had done for many at Loring AFB.

The longer hours were perceived as a particularly heavy burden by NAS Alameda workers since, whether justified or not, there was a strong feeling among them that shore duty is supposed to some extent be a relatively easy tour to compensate for their rigorous hours while on sea duty. Whether this perception accurately reflects policy is irrelevant as long as workers feel they are entitled to a "relaxed" tour. As a result of these responses, the working hours at NAS Alameda are currently being altered.

Table 4. Percent of Post-Test Workers Preferring the New (BAS or CASH/A La Carte) or Old (Traditional) Food Service System

	Loring AFB (N = 30)	NAS Alameda (N = 25)
Extremely prefer new system	63%	24%
Slightly prefer new system	20%	12%
No preference	10%	8%
Slightly prefer old system	7%	28%
Extremely prefer old system	0%	28%

Conclusions

1. The inclusion of worker opinion in the evaluation of food service systems is an important contribution to the overall assessment.

2. An instrument like the Job Description Index, although designed to evaluate individual worker job satisfaction, can be a useful tool in evaluating worker acceptance of a system.

3. A food service system which satisfies customers and workers in one military service or at one military base is not necessarily acceptable in another service or at another installation.

4. The job satisfaction of military food service workers, as measured by the Job Description Index, is lower than that of the average, civilian non-food service worker.

References

Branch, L.G., Meiselman, H.L., and Symington, L.E. A consumer evaluation of Air Force food service. U.S. Army Natick Laboratories Technical Report 75-22-FSL, 1974.

Smith, P.C., Kendall, L.M., and Hulin, C.L. The measurement of satisfaction in work and retirement. Chicago: Rand McNally & Co., 1969.

Symington, L.E. and Meiselman, H.L. The food service worker and the Travis Air Force Base experimental food system: Worker opinions and job satisfaction. U.S. Army Natick Development Center Technical Report 75-94-FSL, 1975a.

Symington, L.E. and Meiselman, H.L. Job satisfaction and opinion of the Air Force food service worker. U.S. Army Natick Research and Development Command Technical Report 76-11-FSL, 1975b.

ORGANIZATIONAL EFFECTIVENESS RELATED TO NON-MOS TASK PERFORMANCE OF A CRITICAL SUBGROUP

by

Raymond O. Waldkoetter, Ed.D.
Arthur C. F. Gilbert, Ph.D.

Individual Training and Performance Research Laboratory
U.S. Army Research Institute for the Behavioral and Social Sciences
Arlington, Virginia 22209

Helmut H. Hawkins
Roy B. Wethy

Plans and Occupational Development Division
U.S. Army Military Personnel Center
Alexandria, Virginia 22331

A Paper Prepared for Presentation at the
18th Annual Military Testing Association (MTA) Conference
October 18-22, 1976

ORGANIZATIONAL EFFECTIVENESS RELATED TO NON-MOS TASK PERFORMANCE OF A CRITICAL SUBGROUP

Raymond O. Waldkoetter, Ed.D.
Arthur C. F. Gilbert, Ph.D.

U.S. Army Research Institute for the Behavioral and Social Sciences[1]
Arlington, Virginia 22209

Helmut H. Hawkins
Roy B. Wethy

Plans and Occupational Development Division
U.S. Army Military Personnel Center
Alexandria, Virginia 22331

When there is an attempt to measure organizational effectiveness the effort must become one of selecting those criteria that provide some observational feedback to evaluate whether an organization achieves enough of its objectives or functions reliably enough to maintain an acceptable performance level (Bowser, 1975). Complex measures of effectiveness are being explored, believing that eventually the right combination will yield a workable index or indexes for states of effectiveness. Several definitions of important evaluation critera emphasize the value of fulfilling operational commitments and are clearly performance oriented.

In this paper our aim is to discuss organizational effectiveness from the point of view that mission capability may well be the final criterion. Unit performance accordingly under field operational conditions requires tasks which can impair unit effectiveness, in spite of qualified personnel in specific jobs and with proper equipment, unless such tasks are performed as the underlying support to sustain the continuity of the intended unit work duties. The unit mission capability rests then on whether a large range of tasks are accomplished which extends beyond the primary duty assignment of any given individual. Our further aim in this paper is to submit both an allegorical narrative and a simulated rational analysis of manpower authorization and task inventory data, and is meant to illustrate a few of the difficulties in coping with the objective of relating personnel allocation to organizational effectiveness. Arousal of healthy critical responses is likely,

---

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research Institute, the Military Personnel Center, or the Department of the Army.

and perhaps such stimulation will induce later creative replies from those more informed critics to instruct us in the correct analyses for maximizing personnel utility.

There are occasions where one finds the proficiency of personnel is wholly adequate within the boundaries designed for the group of articulated jobs, i.e., those positions in a unit structured for a particular set of duties and tasks, which serve to reach an acceptable level of productivity. However, since a military unit must exist in varying environments, from the installation into deployment sites, it should be obvious that unit, team, or individual members will not always have narrowly defined duties. They will and must perform tasks related to other duties and jobs. These tasks which are termed Non-MOS can be found within other Army Military Occupational Specialties, but continually under the dynamic field situation these are assigned as additional and special tasks in the course of duty requirements. Also the individual must be prepared to facilitate military operations by lending a hand whenever these supportive actions are needed. The multiskilled output of any military member must have attention to appreciate what Non-MOS tasks do to sustain unit operations and how numerous affects are associate with these tasks. Nearly 500 tasks listed across 16 separate duty or functional areas have been identified by the Plans and Occupational Development Division, U.S. Army Military Personnel Center, indicating the diverse range of Non-MOS tasks which may be expected from soldiers, regardless of MOS assignment (Wethy, 1976).

METHOD

Because mission capability in this paper is the assumed final criterion of organizational effectiveness, a critical subgroup (X) will be analyzed in comparison with the majority personnel grouping (Y) to explain the possible consequences if certain high priority Non-MOS tasks are not adequately performed and the resulting impairment to effectiveness follows. Since it is impossible to look at the nearly 500 Non-MOS tasks and possible situations where they can occur, task examples of high priority in two MOS will be addressed with special interpretive comments. High priority here is directly related to a requirement as part of the situation and the pressure to have a task done to avoid immediately negative consequences.

The MOS used for reference are 71C, Stenographer, and 94B, Cook. These are MOS that are often particularly designated as having the most favorable utilization of Type X personnel. The Non-MOS tasks as found in the 71C and 94B inventories administered to Groups X and Y were not designed to probe differences between the groups. Where differences may occur, the distinctive group responses represent the separate group responses based on their own job experiences. It must be understood that when tasks are reported as not performed on a relative time spent scale, there is no categorical indication individuals cannot perform, but rather that they usually do not perform given tasks and may not have

the skilled practice to accomplish these tasks as proficiently as those personnel regularly doing so.

The question should arise as to what will really indicate a noticeable variation of performance in regard to Non-MOS tasks for Group X. In the sense that the personnel in either group may be qualified in their regular MOS duties and tasks, the next concern is at what levels of performance might comparison be made to observe differential behavior in the work environment. The critical levels become the percentages of Group X which do not perform certain tasks and those points where such tasks are not performed enough in the job context to expect that critical duties and tasks will be completed satisfactorily. Also the eventual troublesome situation is seen in moving from the individual to unit organization with the relative amount of task work that is unable to be carried through to a level of satisfactory performance.

Should the normative approach seem justifiable here, it is not illogical to observe, for example, that if Group X does a task 25 percent less often than Y, and there are four X members with eight Y members in a squad the likelihood of having a task proficiently completed without regular job experience begins to diminish. Unless X members have exceptional motivation and some form of current training, would it be a good bet to assume the XY squad will be as productive on the given task as an all Y squad even selected at random? Most replies will tend to reject the XY squad without extra knowledge that a Y squad might be just as limited in a similar task. Relative effectiveness (TRADOC Pamphlet 71-8) might be computed for this task leading to the observation that there was 25 percent task loss across each of four Type X personnel. This means the total loss of task performance at the individual level of summing, equal to at least one person's performance or roughly an effectiveness of .9 for the squad of 12 soldiers. If further task decrements occur in other tasks and interrelated series of tasks, there may be those incidents where complex duty-task behavior does not match the pressure to respond fully to a unit task requirement in spite of compensating use of supplemental actions.

Assuming the estimated available working hours of 3000 hours annually per individual soldier, according to manpower authorization criteria (AR 570-2), and there are 12 people in the squad, then 36,000 hours would be projected for total duty-task performance. When a standard loss rate can exist over that even allowed by the AR cited above (FM 101-10-1), then the loss rate of .005 percent of productive time may equal 180 hours. We must remark here that these computations are not assumed to be highly accurate, but they do err on the conservative side in keeping with a long history of administrative caution. Thus with about 250 hours for a working month per soldier, and taking .004 percent as the amount of task time spent "firing an individual weapon" in MOS 71C (Bonette and Hutton, 1976), then only one hour would accrue for firing performance per month per soldier. Applying the tentative relative effectiveness of .9 which was estimated for this task, about six minutes or 72 minutes across the squad would be lost for the month, or more than

a full firing day annually. This does not seem like much of a loss; however, the old adage "for the loss of a nail, the loss of a shoe, the loss of a horse, the loss of a rider, the loss of a kingdom" must gnaw at collective memories of guidelines for improved personnel resource management.

RESULTS

As a commander or supervisor directs the activities of a group, a motive is unavoidably to maintain the capability to improve the organization's ability to achieve its mission goals (Beckhard, 1969). With a part of any group showing a set of performance limits by either not doing tasks or being trained to perform by differing priorities, the pressure on any organization so disposed will erupt in a series of detectable symptoms. These symptoms appear under conditions where the objectives for the mission goals are not satisfied by learned tasks, responded to with the proper task sequence of behavior, nor attained by adequate proficiency to maintain quality control standards. Task performance that is incomplete or not to be expected places a strain on organizational control, when it is most urgent, and the probability of getting the related objective achieved becomes less favorable. What level of stress might an organization be able to absorb before the task performance deficit was too large or dangerous?

Again, tentatively accepting the "relative effectiveness index" of .9 as a reasonable point at which performance by the imaginary squad will hold on the given task, then little excitement will be shown if a criterion level of .7 would result in getting the job done in a minimally acceptable way. However, the stress is apt to become evident if the squad is not directed to operate as a full working team but must separate, let us say, in to three four-member teams because the task need not always utilize the full squad. Quite suddenly there is induced stress using a team from the squad, because the task now becomes critical in terms of pressure to react but can be done by fewer members working for a longer or more intense duration.

Options for using the three teams will decrease as the relative effectiveness of personnel may decrease, but indexes are held for performance, because there is less expendable reserve for the task behavior at the demand time for performance. Should the task have the minimally acceptable standard of .7 as the immediate capability, then any condition imposing more stress, may even reduce the capability below the minimally acceptable level.

If the X group, for example, is 33 percent of the squad (four soldiers) having the total task inexperience and it is ordered to carry out the given task behavior for several hours as a separate team, doing all the squad usually would do in less than an hour together, then X group cannot finish the assignment working at a given level of relative effectiveness further diminished according to time actually required to

complete in comparison with the normative time usually experienced soldiers are allowed. That is the rate of work is not fixed but accelerated with the less experienced personnel in place so that task elements have to fall faster in a given time sequence to reach a state of completion. Inexperienced personnel will probably not do enough, soon enough, to effect the coordinated task behavior necessary to link tasks in productive action. The intertask problem of going from a similar task to the task being performed will affect how the task performance is carried out if the prior task carried any performance deficit. A series of high risk tasks that are somewhat immediate, demand aggressive or initiating behavior, and have a cyclic need for stress endurance and stamina, may only be completed in a team operation when the weakest member surviving is able to do the most arduous task. The imaginary squad and the X team group will run into work adaptation crises as they strive to respond to a set of stimuli surrounding the task for which they are inexperienced and have not routinely carried out in sequence with other related tasks.

Another source of operational stress comes about also when there is an occasion, or more aptly an emergency, that preempts the work energy and the task sequence must be expanded for each individual. If the X team group is working under a disadvantage while representing 33 percent of the squad, a redistribution of X team strength with Y group members may not be any better, if the critical task falls directly to any X member without any of the Y group in support. A dilemma of sorts arises where the X team operating independently on the given task induces a probable operational organization risk. Then again one team member, or about only eight percent of the squad, might be unacceptable, if the task risk would be equally anticipated in the line of duty wherever an X type squad member was placed. It should be noted here that an inverse relationship is nearly observable. This relationship is: as an organization loses strength in numbers of assigned personnel, then any available individual must have the immediate potential to exceed job standards on one or more critical tasks. It will not impose nor ask any exaggerated imagination to recognize that one marginally trained or utilized individual could impair the mission capability as the loss rate of productive time approaches a traditional warning sign of only 70 percent of a job task being finished.

A cursory extension of the loss rate, in that eight squad Y members may remain to do a job task at the minimally acceptable level, can bring about a dubious assignment of any X group member, or in this situation a replacement, unless there is a highly individualized personnel assignment plan. So if one X team group member assigned can, at slightly over 10 percent of the residual squad, increase a risk for borderline performance and a 33 percent squad strength of X personnel causes a reduction to limited capability with a fairly flexible assignment policy, an inference resulting is that to minimize a threat to the total force the lower eight percent of squad projection should be used in fully operational situations and the higher 10 percent projection in more normal installation and field settings. Therefore, the maximizing 10 percent level may

seem too cautious for some proponents but for assignment practicalities the target number could hit between a projected 70 and 75K for a force like the Army with its fluctuating personnel accessions and losses.

DISCUSSION

Having a critical subgroup of just over 70K personnel in a total force would obviously offer flexibility in allocating more than 10 percent in controlled organizational settings. Based on several trends in occupational analysis data and tendencies to show more variations of performance, Group Y, it is hypothesized, will continue to perform more tasks, and generally more often than Group X, and Group Y will have an obviously greater relationship with supplies, vehicles, and weapons.

A choice of filling Army organizatons with Type X personnel to levels exceeding 10 percent across the force structure could encourage serious consequences that limit effectiveness for key mission requirements. An assumption that Groups X and Y are equal in potential to perform all job tasks in spite of biographical data and normative job experiences will lead to problems at the individual, team, squad, section and unit levels unless comparable reliability on tasks is demonstrated. There should be more quantifiable research in defining the audit track from relative individual effectiveness, to small group and team behavior, and onto the basic unit performance standards. Organizational effectiveness is measurable along probably six to 10 dimensions and there must be a set of procedures by which those dimensions can be combined in an equation to predict several pertinent criteria. A criterion which has a lucid track from the initiating task behavior through to a terminal action or product object must at some point in time and processing movement have a contributory effect of either minimizing or maximizing the intended results.

Returning to the referenced MOS 71C and 94B, and perspective from which Non-MOS tasks are observed, the negative direction of performance differences is seen in Table 1 for the Group X members.

As the task deficit in performance accrues in Non-MOS tasks, operational capability in a field environment cannot be thought of as improved, rather the probability increases of having critical tasks: difficult to learn because of short practice time, having obvious consequences for unit efficiency, and having immediate reaction times. The tendency for Group Y members to perform more tasks will only grow if the mission requirements are relocated from garrison to a dynamic field situation and scenario.

Recalling the 180 hours, cited earlier, as a probable loss to the squad, nothing will guarantee this loss to be equally spaced for each squad member over an annual work schedule. One or more members could have large time losses concurrently. The question then is: If the already less effective shooters as an example, lose firing time, and

TABLE 1

TASKS WITH LARGE DIFFERENCES IN PERCENT MEMBERS PERFORMING: GROUP X VS GROUP Y

Tasks Non-MOS	Group X 71C	Group X 94B	Group Y 71C	Group Y 94B	Difference 71C	Difference 94B
Load Equipment Supplies on Vehicle	3.3	10.7	27.9	28.8	-24.6	-17.1
Attend Weapons Training	9.8	3.5	32.6	28.8	-22.8	-25.3
Attend CBR Training	11.5	14.2	32.6	27.7	-21.1	-13.5
Drive Vehicles	1.8	3.6	30.4	21.3	-28.6	-17.7
Requisition Supplies	14.3	7.1	45.8	23.9	-31.5	-16.8
Non-MOS Task Means for Performance	8.1	7.8	33.8	26.1	-25.7	-18.(3)*
	(7.9)%		(29.9)%		-(22.0)%	

*Possible Error in Averaging ( )

more firing is probably needed to maintain a proficient level due to the natural perishability of that complex skill, from what duties and tasks will the leader or supervisor reapportion his productive time when there is apparently a state of task imbalances already operating to constrain organizational effectiveness? Although a critic of task analysis data can say these data presented are suspect, normative and not truly criterion-referenced against consistently demonstrated measures of effectveness, task inventory survey information (Christal, 1971) has shown a very respectable degree of stability. To dismiss the trend of lessened effectiveness for Type X personnel based on a review of tasks in MOS that have been a traditional location for these personnel, simply must increase the allocation risk if they are sent to MOS demanding more Non-

MOS tasks. Where Group Y has quite often 20 percent more members performing Non-MOS tasks than Group X, and it is not unusual for over 20 percent of duty time to be spent in Non-MOS tasks, Group Y must have a large share of Non-MOS tasks and spend more time with those tasks than Group X.

A target figure of 70 to 75K for Type X personnel is hazardous and to go over this upper boundary would seem to imply a motive to prove that a greater risk is permissible to gain an exceptional employment objective than to insure a fully operational force for the defense of all this nation's valuable human and natural resources. If the above figure is in error, what logical quantitative method is applicable to estimate the effects of individual performance in a unit on mission capability? An earlier and rather expedient effort following an ad hoc Delphi solution using TO&E proponents on an Army-wide survey (TRADOC, 1975), allowed an average of 25 percent Type X personnel for the total force. This would project to about 170K; however, the assignment decision process assumed, in our opinion, that a choice of X or Y personnel was made initially having the Y type pool available or in the unit, but this logic does not systematically recognize the complex of Non-MOS tasks or personnel losses. So a correction factor of replacement should go from either Y or X, to Y if X is not equal to guarantee the allocation of Y personnel. Thus there is not an alternate allocation of Y, if spaces are reserved for only selected X's. The forced-choice replacement to offset the possible X task decrement would suggest at least a 50 percent reduction in the projected 170K to 85K. This ad hoc Delphi solution also did not look like it took into account attrition, lost time, casualties, training, and promotion-rotation assignment constraints which might reduce this figure closer to the 70 to 75K duty position placement of Type X personnel. It must bear noting here that this 70 to 75K range is a gross estimate for the sake of encouraging equitable utilization. The actual range of qualified personnel in the population could hold it to about 60K if aptitude, recruitment, and personnel flow obstacles are undetected and later computed as essential constraints.

While our inferences in arriving at a projected total figure for Group X in the Army may fall short of the disciplined and quantifiable evidence you believe necessary, we shall only reply that we shall be gratified if someone can give us an equitable method for computing strength for manning tables which account for individual and group characteristics. Instead of seeing the Army as an extension of an employment opportunity program, might we not think of it more as an insurance oriented organization and refer to a system of actuarial tables showing the risk for the organization and person and the probable cost if a Type X decision to enlist is made. The assignment model for Type X personnel should insist that basic, AIT, and OJT standards are maintained to realize a level of readiness performance (Benade, 1975). In any of those jobs within given TO&E's where the risk of task decrements is eminent for at least several categories of Non-MOS performance, then assignment test restrictions should exist to admit only the individual by waiver based on proven skills. And now that we are at the end of our paper we confess we did not elaborate much on the traits of Type Y personnel, and

did not want to, because Type X represents a critical subgroup composed totally of female soldiers.

REFERENCES

Analyzing training effectiveness. TRADOC Pam 71-8. Fort Monroe, Virginia: U.S. Army Training and Doctrine Command, 1976.

Army Regulation 570-2, Manpower and equipment control. Washington, D.C.: Headquarters, Department of the Army, July 22, 1969.

Beckhard, R. Organization development: Strategies and models. Reading, Massachusetts: Addison-Wesley Publishing Company, 1969.

Benade, L. E. (Lieutenant General, USA, Retired). Personal Communication, April, 1975.

Bonette, J., and Hutten, R. Occupational survey report for stenographer, MOS 71C. Alexandria, Virginia: U.S. Army Military Personnel Center, 1976.

Bowser, S. Measurement of unit effectiveness: A criterion development problem. In Colonel G. W. Orton (Ed.), Proceeding of the 17th Annual Conference: Military Testing Association. Fort Benjamin Harrison: U.S. Army Enlisted Evaluation Center, 1975.

Christal, R. Stability of consolidated job descriptions based on task inventory survey information. Lackland, AFB, Texas: Personnel Research Division, Air Force Human Resources Laboratory, August, 1971.

Staff officers' field manual. FM 101-10-1. Washington, D.C.: Headquarters, Department of the Army, July 26, 1971, Chapter 4.

WAC content in TOE units. Unpublished staff report, 1975. Fort Monroe, Virginia: U.S. Army Training and Doctrine Command.

Wethy, R. B. Non-MOS task inventory. Unpublished manuscript, 1976.

ABSTRACT

Evaluative Criteria for
Comparing Models of Unit Performance

Donald E. Erwin
Army Research Institute

At the request of the Army Training and Doctrine Command, the Army Research Institute (ARI) reviewed and evaluated eight different models of unit performance. The results of this review were provided to a committee composed of representatives from various training agencies and proponent schools. The mission of this committee was to choose one model for implementation at the operational test of a combat simulation system slated for late 1978.

To aid the decision-making task of the committee, ARI developed a criterial framework to facilitate comparison of the different models. This framework was used in two ways. First, after presenting the models of unit performance to the committee, the purpose of the criterial framework was described and the various criteria and categories were clearly defined. Committee members were then asked to evaluate each of the models using the framework. Variance among committee members' responses, and, between committee responses and ARI responses served to focus further ARI efforts to convey accurate and thorough information about each model to the committee. Lack of variance between ARI evaluations and committee evaluations of the various models was finally achieved and indicated that cogent descriptive information about each model had been successfully conveyed to the decision-making committee. Second, when the evaluation of each model in the criterial framework had been consensually validated, committee members rated the importance of criteria in the framework. With this information, committee members could compare the models on each criterion, and, "weight" their considerations according the rated importance of the different criteria.

The following paper briefly discusses the various models of unit performance and focuses in more detail on the construction of criteria for evaluating such models. Specific attention is paid to the use of such methods as aids to group decision-making. The construction and use of the criterial framework is discussed as an example of evaluation methodology in which the evaluation team provides decision alternatives vis-a-vis user needs and systems by delineating, obtaining, and providing useful parameters for the judgmental process.

Development and Use of Evaluative Criteria for Comparing Models of Unit Performance

INTRODUCTION

Current training and readiness needs of the U.S. Army represent a drastic change from those requirements formerly met by the old Army Training Program (ATP) and the Army Training Tests (ATT). The ATP outlines the minimum essential training for active and reserve units and assumes that "training to readiness" would occur during mobilization. Considering the time frame within which initial hostilities would escalate to theater level operations in the present political and tactical environment in Europe, this approach to readiness is hardly feasible and potentially disastrous.

In 1971, the Board for Dynamic Training (headed by BG Paul Gorman) redefined training needs for the modern Army in a critical analysis of existing training programs. The Army Training and Evaluation Program (ARTEP), implemented in 1974, reflects the recommendations of the Gorman Board. The ARTEP is based on a series of training and evaluation outlines (T&EOs) which specify "...the general conditions under which the mission is to be performed, the primary training/evaluation standard upon which the element will be evaluated, and the performance oriented training objectives which describe the tasks, conditions, and standard for the mission."[1].

The ARTEP, in contrast to ATP/ATT training methodology, provides a framework for collective or unit training. In addition, evaluation of training readiness is structured along guidelines for acceptable individual and unit performance. Because ATTs are generally felt to be inappropriate for peacetime Active Army and Reserve component training and evaluation, ARTEPs are currently being written and validated for various units.

A recent unpublished report submitted to the Army Combat Arms Training Board (CATB) by the Army Research Institute (ARI)[2]. presents the results of an extensive and careful effort to validate the methods and objectives of the ARTEP in light of the guidelines originally provided by the Board for Dynamic Training. In addition to numerous specific suggestions for modifying and improving particular ARTEP procedures, this research led to the following general conclusions concerning the methodological framework of the T&EOs:

---

1. Army Training and Evaluation Program for Armor Battalion and Combined Arms Task Force, ARTEP 17-35, July 1974.

2. ARTEP Validation Report, J.F. Hayes and M.R. Wallis, ARI, March 1975.

(1) Although the missions* specified in the ARTEP represent valid performance criteria for combat readiness, the standards of performance are often "too general and vague", requiring heavy reliance upon the judgment of individual validators amongst which interrater reliability is low.

(2) The present ARTEP is not a standardized test instrument; consequently, there are no legitimate ways to extrapolate from task performance to mission performance.

(3) The ARTEP potential for use as a training analysis tool has not been fully exploited. The feedback from testing to training requires improvement. Since the ARTEP represents the culmination of the training scheme, the analysis of deficiencies in various tasks are not fed back to the units, and the information is essentially wasted.

(4) The present ARTEP is more mission oriented than training oriented. Consequently, the quantity and salience of diagnostic information depends on the predispositions of the evaluators, and often, this information is inadequate or even nonexistent.

The ARTEP has succeeded in meeting some very fundamental training needs, but both ARI and CATB view the present "State-of-the-ARTEP" as being a first generation effort requiring modification and refinement. These modifications and refinements have two foci. First, there are methodological problems, such as those cited above, that have to be solved. Second, the ARTEP, as a training methodology, has to remain responsive to organizational changes, doctrinal guidance, introduction of new major weapon systems, etc. That is, modification and refinement of both the methodology and substance of the ARTEP must follow the evolving training and readiness needs of the Army.

One of the most clearly articulated current readiness requirements is that combat units are combat competent as soon as training is completed. As mentioned earlier, the nature of current threat situations requires decisive and immediate victory at the onset of hostilities and does not allow for the "seasoning" of troops in prolonged conflict. Although the ARTEP provides combat-related missions and tasks embedded in a general evaluative framework, a method for training tactical competence in combat is lacking. Responding to this need, ARI introduced SCOPES (Squad Combat Operations Exercise, Simulated) at the U.S. Army Infantry School at Fort Benning, Georgia, in 1973. SCOPES is a training technique that simulates the casualty producing effect of live-firing in two-sided free-play field exercises. Infantry soldiers use 6X telescopes mounted on their M16s to identify 3-inch numbers on "enemies" helmets. Blank rounds, reported enemy numbers, and, confirmation of "kills" contribute to a high degree of realism and consequent personal involvement on the part of the trainees.[3] REALTRAIN was extended to include armor and

*Author's emphases.

3. REALTRAIN-USAREUR Research Project, R.T. Root, et al., ARI, 1976, in press.

antiarmor weaponry into training exercises by 1974. USAREUR implementation of combined arms REALTRAIN took place in the 2nd and 3rd quarters of 1976 and preliminary FORSCOM implementation is planned for the 1st quarter of 1977.

ARI developed SCOPES and REALTRAIN in order to provide a training method for tactical training that would complement and eventually be integratable with the ARTEP. In SCOPES and REALTRAIN exercises, the individual and unit faces conditions and needs isomorphic with those found in real combat, rather than the conditions and needs peculiar to the training procedure. The research direction of ARI in this area has been guided by "...the principle that training and evaluation must be directly related to the individual's job and the unit's mission."[4]. Emphasis has been placed on studying the "...individual's job and the unit's mission in a systematic manner in a realistic setting."[5]. This setting is, of course, combat and only a part of ARTEP T&EOs are directly related to combat and require SCOPES or REALTRAIN methodology. The Infantry Battalion ARTEP presently makes use of SCOPES methodology for that portion of its combat T&EOs where units make contact with the enemy. As of the writing of this report, combined arms REALTRAIN has not been integrated with any ARTEP.

In addition to developing training techniques that are designed to support combat-related T&EOs in various ARTEPS, ARI has conducted research on refined methods of evaluating unit performance. With contractual help from the Litton-Mellonics Corporation, ARI has investigated possible combat unit evaluation systems (CUES) to determine what evaluation procedures yield the most cost-effective assessment of training readiness as well as useful, pertinent information to both trainer and trainee. The CUES project determined what criteria combat veterans and experienced trainers at the Armor and Infantry schools use in evaluating training procedures. Given that various evaluation schemes exist which can assess the results of training procedures at a number of levels and foci, the evaluation criteria endorsed by military training experts can be used to determine which evaluation alternatives represent the most pertinent approach for collecting salient information about combat unit performance.

The two foci of ARI's research in developing an approach for achieving readiness of combat units have been, then, the design of training techniques that simulate combat and foster learning of tactical skills, and, the consideration of various combat unit evaluation systems for gauging training efficacy. Both areas of endeavor are envisioned as complementing the ARTEP, and, as being integratable with the ARTEP.

---

4. Development of a test bed for evaluation of small unit doctrinal alternatives in the combat arms, R.T. Root and L. E. Word, ARI, 1974. (AD 785 668)

5. Humar performance, jobs, and systems psychology, J.E. Uhlaner, BESRL Technical Report S-2, 1970.

A recent in-house effort at ARI in the Unit Training and Evaluation Systems Technical Area has brought research on training and evaluation to bear on one concern: the Multiple Integrated Laser Engagement System (MILES)** second operational test (OT II). CATB, looking ahead to the second operational test of the MILES hardware, has asked for the collaboration of ARI in developing a supporting training and evaluation model couched in a modified and refined ARTEP format. ARI has been tasked to modify and refine existing ARTEP methodology so that it can best support tactical training with MILES hardware, and, best provide an evaluation of training readiness achieved with MILES. Designing an engagement simulation ARTEP has focused primarily on modifying and refining the ARTEP evaluation methodology. Since the training aspect of the ARTEP is dictated largely by the fact that the exercise is a free-play engagement, there are considerably fewer degrees of freedom in designing how training will proceed than in determining what will be evaluated and how. As a point of departure for modifying and refining ARTEP evaluation methodology, ARI has drawn on the criticism offered in the ARTEP Validation Report, the information acquired in the development of various CUES alternatives, and, principles of educational evaluation developed in non-military settings.

The ARTEP Validation Report had specified that performance standards specified by T&EO outlines in the ARTEP are often general and vague, (and subjectively applied), that the ARTEP's potential for use as a training analysis tool was not fully exploited, and, that the ARTEP was more mission- than training-oriented consequently deemphasizing diagnostic information. Any alternative formulations of an evaluation scheme would have to reflect these criticisms. In addition to the guidelines offered by the ARTEP Validation Report, it was felt that an important consideration in developing an ARTEP modification should be the nature of the data that would be evaluated and the conditions under which it would be collected. Rigorous application of the scientific method in evaluating the effect of training procedures was inappropriate. Testing the effect of varying a given parameter while holding others constant was rightly concluded to be wholly impossible in the context of the in situ assessment necessary for evaluating training procedures.

Consequently, various guidelines used in education evaluation were adopted in developing alternative methodologies for assessing unit performance. Specifically, it was concluded that (1) intervening variables must be so arranged and controlled that they become an integral, organized part of the data, unlike the laboratory situation where the effect of intervening variables is hopefully eliminated, or, randomized across the design. (2) Terminal performance or final results should not be the sole issue of the evaluation/assessment such that the testing environment has to be controlled or artificially contrived in order to focus on the final effect of a small set of variables. Instead, the testing environment should facilitate the real-time collection of diagnostic information.

**A second generation engagement simulation system in which low power, weapon-mounted lasers activate sensors on "enemy" personnel and/or vehicles in free-play field training exercises.

(3) Expanded decision criteria for results must be used instead of a "go/no go" conclusion on the basis of certain parameters. (4) Treatments/parameters must be amenable to refinement and modification during the evaluation such that the training effort may benefit during, and not solely after, evaluation and subsequent recommendations regarding remedial training.[6].

On the basis of these guidelines from the ARTEP Validation Report and methodological approaches from educational evaluation, ARI proceeded to construct and consider eight modifications of the ARTEP evaluation methodology. Four of the candidate models represent a modification of existing ARTEP evaluation methodology in that they rely on information collected during the exercise for assessing unit performance, and can be termed General ARTEP Models. The other four models are concerned with final mission outcomes as the bases for evaluation and can be called Terminal Product Models.

Table 1 shows the characteristics of the alternative evaluation

---

Table 1 about here.

---

methods. For each General ARTEP and Terminal Product Model, general characteristics are listed as well as information regarding what the model yields in terms of (a) a determination of unit readiness, (b) subunit readiness, and, (c) diagnostic information useful for prescribing remedial training.

General ARTEP Models:

(a) The ARTEP "As Is": The ARTEP methodology for evaluating unit performance would provide limited opportunity for engagement simulation (i.e. MILES) scenarios due to a general deficiency of company/team and battalion/task force combat missions specifications. Evaluation, as outlined earlier, is based on subjective judgements regarding the quality of performance observed.

(b) ARTEP with Engagement Simulation (ARTEP/ES): In this model, provision for engagement simulation scenarios at all echelons is included, although the actual mode of evaluation would remain the same as that in the ARTEP "As Is". The T&EOs within the ARTEP would overlay engagement simulation in the same fashion that SCOPES is included in the Infantry ARTEP for rifle squad evaluation.

---

6. Educational Evaluation, PDK National Study - Committee on Evaluation, F.E. Peacock Publishers: Itasca IL, 1971.

(c) Situational Dependent ARTEP (SIT DEP): The SIT DEP model would be comprised of modified T&EOs to collect objective data during all phases of a mission. The evaluation would be based on performance standards synthesized from the dynamics of the on-going scenario, rather than on external standards. The evaluation would be flexible enough to allow situational variables to contribute to performance standards and would allow an analysis of the role intermediate "engagement" outcomes play in mission accomplishment.

(d) Objective Based ARTEP (OBJBAS): This candidate ARTEP evaluation model would require an exhaustive list of detailed objectives developed along the lines of FM 21-6, How to Prepare and Conduct Military Training. The evaluation would be focused on meeting the standards specified via the objectives. The objectives would deal with the performance of sub-elements and individuals and be developed through task analysis and mission analysis procedures (much as the reconnaissance patrol example in FM 21-6).

Terminal Performance Models:

(a) Unit Performance Assessment Model (UPAM): In this model, standards of performance designated as acceptable would be set by the local commander according to what was considered to be "minimum acceptable achievement" and "maximum acceptable cost." The cost and achievement dimensions refer to the end result of an engagement exercise and represent aggregations of data from various sources such as casualty counts, enemy weapons captured, time required to reach objective, etc. The evaluation consists of determining if the actual outcome is acceptable according to the standards (earlier designated by the unit commander) via mathematical technique. The evaluation is wholly objective in nature.

(b) Unit Performance Assessment Model/Fixed Standards (UPAM/F): This model is the same as UPAM except that the standards for acceptable performance would be set by higher headquarters.

(c) Terminal Objectives ARTEP (TO): The terminal objective model would set standards explicitly in terms of mission accomplishment, time taken, friendly casualties allowable, and enemy casualties required. The evaluation would be based on an objective assessment of exercise products.

(d) Terminal Objective ARTEP with Diagnosis (TOD): The standards for this model would be the same as the previous model. The essential difference is that diagnostic information would be gathered during the exercise. That is, information describing how a unit conducted its mission (e.g., movement techniques, troop leading procedures, content of tactical communications, etc.) would then be utilized for training feedback. This information would not be used, however, in the evaluative process.

A more parsimonious comparison of the eight ARTEP evaluation models is shown in Figure A.

Figure A about here.

The eight models are shown to vary along two dimensions: PROCESS vs. PRODUCT information as the main contribution to OUTPUTS, and, READINESS vs. DIAGNOSTIC information as the OUTPUT of the evaluation procedure. Figure A shows each of the models as a function of these two dimensions. For example, the OBJBAS model (no. 4) yields readiness and diagnostic information as its outputs. These outputs come from a combination of end products (EP) plus intermediate products (IP) plus process data (P), all of which are objectively acquired (EPO, IPO, PO) since objectives are designated by an external source and are either met or not met. This combination is represented in the METHOD column as Ev (evaluation) = (EP=EPO)+(IP=IPO)+(P=PO).

PROCESS data is information that describes or summarized how a unit conducts its mission. Examples of PROCESS data include movement to contact, the content of tactical communications, etc. PRODUCT data, on the other hand, is information that describes the results of an engagement or exercise. PRODUCT data is usually quantifiable as casualty counts, final positions, etc. INTERMEDIATE PRODUCTS refer to specific objectives embedded in the exercise, such as acquisition of a specific target, movement to a particular location within a set time frame, etc.; i.e., objectives or action "products" that contribute to final results.

The DIAGNOSIS vs. READINESS dimension addresses the question of whether the evaluative model provides diagnostic, analytic information about the effect of training on individual and squad/unit performance, or, synthesizes a final statement concerning combat and/or training readiness.

Each of the eight models represents a unique combination of not only PROCESS/PRODUCT information processing for input and DIAGNOSIS/READINESS output but also reflect the recommendations of the Board for Dynamic Training, the ARTEP Validation Report, and methodological alternatives from educational evaluation. The alternative models, developed by the research team responding to the eventual training and evaluation needs of the MILES OT II, are collections of principles and guidelines for developing an evaluation model. However, prior to actually developing one or more prototype evaluation methodologies based on one or more of the eight models, the research team was faced with the problem of determining which ARTEP modification constituted the most desirable combination of various characteristics deemed important to the joint purposes of training and evaluation. In a sense, then, the research effort had come to a point where it was time to "evaluate" the evaluation models, and a method was needed. During the development of these alternatives, an effort had been made to formulate criteria upon which the various

modifications could eventually be compared. These criteria were designed to reflect the guidelines followed in constructing the various modifications as well as more practical considerations such as implementability, availability, etc. For example, the OBJBAS model requires a comprehensive and exhaustive set of objectives for all aspects of performance in the scenarios used for training with engagement simulation. These objectives would be established by an authority above the unit commander, perhaps at the level of the Division or even DA level. Given that such an evaluation model would provide objectively gathered information about behavioral processes, and, intermediate and terminal products, it is obvious that a wealth of diagnostic and readiness information would be available. But as an evaluation model that might be used or referenced Army-wide, how implementable is it, how acceptable to the user, will the objectives be available as they are needed for model modification and refinement? These practical considerations, in addition to those derived from the recommendations of the ARTEP Validation Report, the Board for Dynamic Training, and the literature on educational evaluation, provided the source of criteria for building a framework within which the various evaluation models could be evaluated.

In drawing on this source of criteria for building the evaluation framework, the research team decided that resultant product would have to fulfill three standards. First, the criterial framework for evaluating the models would have to allow comparisons to be made in the context of the system or subsystems in which, or with which, the models would be implemented, or, have impact on. Providing a unique information package for each model focusing on the idiosyncrasies of that particular evaluation method would hardly be useful in the decision-making process used to select a model. Since the contingencies involved in implementing an alternative should ideally be a part in any decision-making process, the criterial framework would have to allow for contextual consideration of the various models vis-a-vis relevant systems or subsystems. Second, the framework should delineate relevant information about each alternative via explicit, well-defined criteria that facilitate comparisons among alternatives. That is, criteria applicable to only one or two alternatives would be invalid since they would fail to deal with the full range of alternatives facing the decision-maker for whom the framework is designed to be a decision-making aid. Third, the criteria making up the framework should be isomorphic with the literature and problem area that was responsible for generating the eight models. That is, taxonomies of criteria available from the educational literature were really not appropriate for consideration. The training and evaluation needs of the Army are unique and for the purpose of comparing models of unit evaluation, a unique criterial framework tailored to the Army problem would have to be constructed. Application of educational evaluation methodology to Army problems was felt to be a legitimate endeavor but not the application of those techniques or guidelines that were specific to educational subject matter, such as readily available taxonomies of criteria.

The identification of criteria has been considered in the field of educational evaluation to be of the "...utmost importance, for (the) relevance of collected information to the decision rests upon whether or not the evaluator has in fact hit upon the criteria that are used (in making the decision)."[7].

Thus, the task at hand was to develop criteria for comparing models of unit performance that (1) provided information about the impact of each alternative on related systems or subsystems; (2) allowed comparisons to be made across alternatives; and, (3) reflected the value system of the decision-maker(s). By developing criteria on these bases, and evaluating each of the proposed evaluation models in a common criterial framework, ARI was able to provide pertinent comparative information to each of the decision-makers responsible for choosing one model for development and implementation in the MILES OT II. The provision of this information was expected to not only expedite the decision-making process but also to systematize what could possibly have been an informal, somewhat entropic discussion of the decision alternatives.

The procedure followed in developing these criteria, constructing a criterial framework to aid in the decision-making process, and the actual application of the criterial framework to the problem of supplying information about decision alternatives to the decision-makers is described in the following section.

METHODS/PROCEDURES

The first stage in the development of appropriate criteria for evaluating the alternative evaluation models was for each member of the evaluation team to become a specialist in and "proponent" of one or more of the proposed evaluation models. In this way, expert and thorough information about each decision alternative was readily available.

Group discussions were then held to formulate candidate criteria. One important concern of the team was that each criterion be equally applicable to all the decision alternatives. Several sets of criteria were constructed in the course of this process in addition to those supplied by contractual assistance in earlier CUES efforts. A final criterial framework was developed by drawing on criteria developed in the CUES efforts, and, upon the gradual evolution of criteria that seemed to pertain to all of the evaluation models and which appeared to be capable of yielding the information most crucial and of maximum interest to the military decision maker.

The second stage in the construction of a criterial framework consisted of four members of the evaluation team independently rating each decision alternative on the basis of each criterion. For some criteria, a yes/no answer was appropriate but for others, the determination required a Hi/Med/Lo judgment. The four raters then compared notes on how their determination on the 8 alternatives agreed for the 21 criteria making up the final framework. In this way, the criterial framework could be

7. Op. Cit. 6.

consensually validated: since the four raters were all equally familiar with the 8 alternatives, if the 21 criteria could be understood, agreement should be perfect if the criteria are unambiguous, pertinent, and applicable to all alternatives.

The third stage of this process involved a briefing of CATB and allied training agencies (the decision-makers) as to the nature of the 8 evaluation models and the 21 criteria developed by ARI. The decision-makers then proceeded to rate each of the 8 ARTEP modifications on the 21 criteria, as had the ARI evaluation team. In addition to rating the modifications, the representatives also rated the criteria on a scale 0-5 ranging from "irrelevant" to "critically important" as far as the mission of their particular agency was concerned. Subsequent to the return of these ratings of the 8 ARTEP modifications on the 21 criteria, and, the 21 criteria on the importance scale, efforts were made to isolate "pockets" of disagreement and focus definitional efforts where needed via a briefing session.

The three stages followed in the development and application of the criterial framework permitted the validation of the criteria as relevant dimensions upon which the selection of the most desirable evaluation model could be based. The first stage was principally concerned with defining the various ARTEP modifications sufficiently clearly so that legitimate criteria could be developed and applied across different alternatives such that each criterion was applicable to each alternative. Once the candidate evaluation models were sufficiently well-defined, a criterial framework would be constructed that tried to take into account not only the characteristics of the different alternatives but also the system impact of each alternative and the extent to which the criteria were isomorphic with the values or norms of the eventual decision-maker and user.

The second phase sought to consensually validate the criteria. That is, if serious discrepancies in opinion as to the application of criteria to the various alternatives occured in the ratings by the four ARI personnel, then the criteria were not uniformly applicable.

The third phase was designed to determine if criteria and models were well-defined to the military decision-maker, and, to what extent different criteria differed in perceived importance. This third step represented the most important aspect of the procedure: At the interface of the information provider/definer (ARI evaluation team) and the user/decision-maker some method had to be used to determine if the information provided was relevant and valuable to the decision-maker, and, if the information was adequately and efficiently transmitted to the decision-maker. Substantial disagreement between CATB/representatives and ARI ratings of the various evaluation models would indicate that either criteria or models were poorly defined. Consistent disagreement about any candidate evaluation model, for example, would require renewed efforts on ARI's part to make the definition of that model explicit and complete so as

to provide adequate information about that decision alternative. "Pockets" of disagreement, then, would serve to focus definitional efforts of ARI where needed. Ratings as to importance of the various criteria would serve to prioritize these efforts and also to indicate the extent to which the criterial framework reflected the value system used by the decision-makers.

RESULTS

Stage 1: Figure B shows a matrix in which the 21 criteria are listed for the 8 candidate models.

Figure B about here.

The 21 criteria are categorized under four headings: Database, Validity, Implementability, and, Availability. These four categories derive from the three considerations outlined earlier for utilizing recommendations from the Board for Dynamic Training, the ARTEP Validation Report, and the literature on educational evaluation. Database criteria facilitate cross alternative comparisons on the basis of what data are processed as inputs yield evaluational outputs of various types (e.g., subunit data, process data, cause/effect information, etc.) in the various evaluation models. Validity criteria determine the congruence of what the evaluation models evaluate with the Army's specific needs regarding assessment of unit performance. That is, are the passing requirements clear, is the evaluation procedure objective, etc. These criteria address those aspects of evaluation models with which the Army is concerned, and which are usually not covered in taxonomies of criteria found in the educational evaluation literature. Implementability and Availability criteria seek to delineate the system impact of each model. These criteria consider the personnel involved in implementing each model, whether or not an aggressor unit will have to be trained, etc. Each evaluation model has a unique impact profile in terms of implementability which is dealt with by these criteria. The impact is gauged in the context of systems and subsystems that are relevant to Army operations such as flexibility in the face of varying terrain and equipment, or, acceptability to the military trainer, for example.

The four categories, then, organize the criteria according to the guidelines decided on by the research team for developing a criterial framework that (1) allowed cross-alternative comparisons; (2) gauged system/subsystem impact; and, (3) was tailored to Army needs.

Specific definitions for the criteria in each category are as follows:

DEFINITION FOR THE CRITERIA IN EACH CATEGORY

The criteria ask the following questions about each evaluative framework:

Database

(1) Subunit data used in evaluation (yes/no):

Is information about the performance of each subunit used in the evaluation?

Subunit data available from evaluation (yes/no):

Is information about the performance of each subunit available from, but not necessarily used in, the evaluation?

(2)-(4) Follow the same distinction between "used in" and "available from".

Process data (2) is defined as information that describes how a unit conducted its mission.

Cause/effect information (3) is defined as data collected which indicates why a particular mission or the exercise was a success or failure.

Cost vs. achievement information (4) is defined as data which outlines the costs (dollar and non-dollar) for a given level of achievement.

(5) Casualty assessment system (yes/no):

Is a method for describing number of casualties available in the evaluative framework?

Validity

(6) Objectivity of scoring strategy (Hi/Med/Lo):

How objective are the various decisions as to mission/exercise outcomes (i.e., do evaluators make judgments on what they have seen, or, are casualty tallies, etc., tabulated and considered along specific guidelines)?

(7) Objectivity of product data (Hi/Med/Lo):

How objective is the data collected in the evaluative framework that will be used to determine mission and/or exercise outcome (i.e., does the data consist of evaluators' opinions, or, counting casualties, disabled vehicles, etc.)?

(8) Standards (Flexible/Fixed):

Are the standards for mission and/or exercise success adaptable to each situation or are they fixed?

(9) "How to perform" information provided (yes/no):

Does the evaluative framework provide information about what is considered optimum performance?

(10) Clarity of passing requirement (Hi/Med/Lo):

Is it clear to the participants in the exercise what they have to do to pass according to the evaluation standards?

(11) Gamesmanship potential (Hi/Med/Lo):

How easy is it for unit personnel to determine on what they are being evaluated so that effort is maximal for those tasks and minimal in the balance of the exercise.

(12) Ease of interpretation (Hi/Med/Lo):

Does the output of the evaluative framework provide information that is easy to understand and that is directly relatable to the needs of the user?

Implementability

(13) & (14) (# of Controllers and Evaluators:)

Requirements for support and control personnel for one side of a company/team exercise.

(15) Evaluator training (Hi/Med/Lo):

How much training and experience do the evaluators need to collect data correctly and make decisions according to guidelines in the evaluative framework?

(16) Controlled aggressor environment (yes/no):

Does the opposition have to be fixed in order to employ the evaluative framework properly?

(17) Flexibility (Hi/Med/Lo):

How easy is it to use the evaluative framework with different terrain conditions, under unusual weather situations, when unit strength varies, when quality of opposition varies, etc.?

(18) Ease of employment (Hi/Med/Lo):

Is the evaluative framework easy to use, or, does it involve complicated manuals, guidelines, etc.?

(19) Acceptability to the military trainer (Hi/Med/Lo):

Does the assessment provided by the evaluative framework (or the data collected) provide useful information to the military trainer?

(20) Acceptability to the military evaluator (Hi/Med/Lo):

Does the evaluative framework provide a straightforward and lucid statement about training and/or combat readiness?

Availability

(21) When could the evaluative framework be ready for use?

Now
Soon (6 months - 1 year)
OT II-1 (1 year - Operational Test II)
OT II (approx. Operational Test II)
OT II+ (some point after OT II)

In summary, the first stage of the procedure for developing and applying a criterial framework for evaluating the different evaluation models required drawing on three sources of criticism regarding modification and refinement of the ARTEP evaluation model. However, drawing on these recommendations was to be done according to these guidelines: (1) allow cross-alternative comparisons; (2) gauge system/subsystem impact; and, 3) tailor the framework to Army requirements. Following this procedure, the evaluation team produced four categories covering 21 criteria. The results of the next stage, that of consensual validation, are described below.

Stage 2: The second stage of the construction of a criterial framework consisted of the four members of the evaluation team rating each alternative on the basis of the 21 criteria. The matrix shown in Figure B was used by each member of the evaluation team.

There was complete agreement in 45% of the possible cells and 75% agreement (3 out of 4 raters) in an additional 33% of the cells. Thus, a total of 78% of the possible cells had agreement from at least 3 out of the 4 raters. It was found that in most cases, a misunderstanding of the various dimensions or structure of a proposed model had caused the odd score, rather than interdependence of or lack of clarity in the criteria. Figure C shows those cells in the matrix where disagreement occurred. These disagreements were resolved through further discussion and clarification of the definitions of the models

Figure C about here.

provided by the model's "proponents" in the evaluation team. The final ratings for each model agreed upon by the evaluation team are shown in Figure D.

Figure D about here.

Figure D was then taken to be the reference point, so to speak, for information to be transmitted to the decision-makers who would make the final decision about which model should be developed.

Stage 3: In a conference setting, ARI presented the various models to the decision-makers. After briefing the conferees on the proposed ARTEP evaluation models, copies of the criterial framework were circulated as well as a list of the definitions for the 21 criteria. Since the purpose of the ARI/decision-maker interface was to provide communication of proposed evaluation models from ARI to the decision-makers, and, for the decision-makers to transmit as Army users what they needed in terms of a model of unit performance to ARI, it became critical to insure that ARI had successfully communicated salient, lucid descriptions of the decision alternatives. Consequently, the conferees were asked to fill in the criterial framework on the basis of their perception of the various models, and, to indicate on an importance scale, ranging from "irrelevant-(0)" to "critically important-(5)" how they interpreted the relevance of each criterion to the mission of their agency.

The results of the conferees' use of the criterial framework were used to validate the transmission of information from ARI to the conferees regarding each decision alternative. That is, unless unusual amounts of variance between conferees' perceptions of the models and ARI's ratings were found, it could be assumed that decision alternatives had been clearly stated to the decision-makers. In that the criterial framework was also to be used as a vehicle for comparing the different models by the decision-makers, importance ratings of the various criteria by the users would indicate to ARI if comparative information had been provided, in the form of the criterial framework, along the lines of the users' value systems and requirements. If users decided that most criteria were of marginal importance, then a vehicle for providing comparative information to the decision-makers had been constructed which failed to reflect the value systems of the users, and, which would prove to be of little real help in the actual decision-making process.

Figures E_1 to E_4 present comparisons of decision-makers responses to ARI's on the criterial framework. There was substantial

---

(Figures E_1-E_4 about here.)

---

agreement in all four cases. In fact, considering responses of the four decision-makers collectively, there was 97% agreement over all criteria for the 8 proposed evaluation models. It was concluded that the evaluation team had supplied evaluative information regarding the 8 proposed evaluation models in an effective manner to the decision-makers.

Importance ratings on the criteria were scaled as follows:

0 irrelevant

1 not very important

2 fairly important

3 important

4 very important

5 critically important.

Fourteen percent of the ratings were at 2 or below, and, criteria rated as "not very important" or less were not the same across respondents indicating that the importance of criteria is specific to the needs and interests of the particular school or agency involved. For no criterion was the average importance rating less than "2" (i.e., "fairly important"). The importance ratings of the respondents are shown in Figure F. These ratings served to assist the potential decision-makers in determining areas of disagreement among their own values systems. This point is crucial if the decision is made by a group of agencies with differing training and evaluation needs. The various proposed evaluation models may be perfectly clear but there may be disagreement among the value systems of the various decision-makers which can impede an efficient and expedient use of the provided information. Therefore, differences in rated importance of the various criteria were isolated for further discussion in a final meeting

---

Figure F about here.

---

of the decision-makers with ARI. Written definitions of each criterion

were provided to each conference participant in the original meeting and disagreement in Stage III is presumed to be due to legitimate differences in opinion and not definitional problems.

In a final meeting of the decision-makers with ARI, each decision-maker was provided with a copy of Figure F, and, with a copy of Figure D (the ARI research team's rating of the 8 models on the 21 criteria). The decision-makers were also told that disagreement between their ratings and ARI's was very low and that the decision alternatives appeared to be clearly understood. Using ARI ratings, then, as a reference point (very much like their own), the decision-makers had available a framework comparing the 8 models on 21 criteria. Also available were the rated importances of the 21 criteria by the decision-makers. Thus, comparative information from the criterial framework could be more or less heavily weighted depending on the perceived importance of the particular criterion. Variation in perceived importance was also available to the decision-makers through Figure F.

With these assets, a decision concerning which evaluation model should be developed was made. The SITDEP model was chosen for implementation in the second operational test of the MILES system. ARI had provided the decision-makers with the maximum information available concerning each ARTEP model (i.e., decision alternative), and, the differences between the models, and the importance of these differences. The alternative chosen represents a decision made after utilizing all available pertinent information, and, would therefore tend to be considered the best decision possible.

Following a structured approach to choosing one alternative has several merits. First, although the procedure involves commitment of personnel, the approach is usually more cost-effective than when a decision is made haphazardly. That is, since considerable effort would be devoted to implementing the decision made in the present example, it would be somewhat disastrous to have the decision-makers come to the conclusion, after a decision had been made, that the wrong alternative had been chosen. This conslusion usually is the result of information coming to light after the decision, or, some implication neglected when considering alternatives. Consequently, an exhaustive review of all relevant information and implications tends to minimize the probability of post-decision dissatisfaction. Second, should post-decision dissatisfaction occur, the decision-making process has been documented with a structured, evaluation approach. Thus, criticisms can be validated and changes in action justified if certain information is shown to not have been available at the time an alternative was chosen.

CONCLUSIONS

Before the decision-maker(s) can yield a "choice", three inputs to the decision-making process must be provided to the decision-maker(s) by the evaluation team (ie., the "information provider"). These inputs are the

decision options, the values or needs of the users, and, information concerning contingencies of the various options and congruencies with existing systems. The evaluator can be throught of "...as the provider of these inputs via a process of delineating, obtaining, and providing useful information for judging decision alternatives".[8]. The process can be a sequence of logical steps such that each activity precedes the next according to a plan or method. Delineating is defined as identifying relevant information using a comparison of decision alternatives on the basis of criteria. This delineation of decision alternatives implies an interaction of decision-makers and evaluator aimed at honing the definition of alternatives. Providing consists of filtering information through criteria that best serve the purposes of evaluation and reporting information to the decision-maker.

The inputs to the decision-maker and the process of providing inputs described above were followed by the Unit Training and Evaluation Systems Technical Area of ARI in assisting CATB with the development and consideration of evaluation models for a refined ARTEP tailored to meet the training and evaluation needs of the MILES OT II.

The process followed by ARI consisted of Stages I-III. Each followed logically from the other. Stage I consisted of providing options to the eventual decision-makers: Eight modifications of the ARTEP were proposed that reflected the training and readiness needs of the Army. The different modifications were, of course, congruent with different approaches present within current thinking on training and evaluation needs. In Stage II, ARI sought to accumulate relevant information about the different alternatives by delineating how each alternative was rated according to a set of criteria. These criteria were derived from a consideration of the systems or subsystems that would possibly be affected by implementation of the decision alternative. When the ARI research team reached agreement about how each alternative stood in the criterial framework, the decision-makers were briefed as to the nature of the eight proposed evaluation models at a conference. In Stage III, conference participants rated each proposed model according to the criteria, and, rated the criteria as to their importance for the mission of his agency. Thus, ARI was able to define the values or norms of the decision-makers to determine whether or not the information it had provided about the proposed evaluation models was relevant. ARI was also able to determine whether or not it had successfully and accurately defined each alternative by determining whether or not the decision-makers perceived each model in the same way as ARI. This determination allowed the isolation of pockets of disagreement for further efforts at defining decision alternatives in a final conference.

In a final conference, the decision-makers were provided with the following assets before making their decision: (1) the average perceived importance of each criterion; (2) the 8 alternatives evaluated in a framework of 21 criteria; and, (3) documentation of where disagreement among decision-makers had occurred in importance ratings, and/or, in the

---

8. Op. Cit. 6.

criterial framework. Thus, the decision-makers had been briefed on each decision alternative, knew the differences between the 8 alternatives, the perceived importance of these differences, and, how much disagreement there was among the decision-makers on perceived importance of differences.

In summary, Figure G presents a diagram illustrating the structure of the approach for evaluating decision alternatives as administered by ARI for the choosing of an evaluation model for the second operational test of the MILES training system.

Figure G about here.

	GENERAL ARTEP MODELS				TERMINAL PRODUCT MODELS			
	ARTEP As Is	ARTEP/ES With Engagement Simulation at all Echelons	SIT DEP With Objective Data Elements and Situational Flexibility	OBJBAS ARTEP with Intermediate Performance Objectives	UPAM	UPAM/F	TO	TOD
GENERAL	. Limited engagement simulation . Evaluation based solely on expert judgment . Permits a focus on process information	. Similar to ARTEP in terms of expert judgment . Adds realistic product information	. Process information is included in evaluation . Based on objective data elements	. Evaluation based on exhaustive list of detailed objectives . Objectives developed for both process and product information	. Evaluation based on mathematical trade-offs between cost and achievements . Product information is the basis for score . Cost and achievement standards set by local commander	. Similar to UPAM but has fixed cost and achievement standards	. Evaluation based on unit objective by mission . Standards explicit and based on products	. Similar to TO but includes process information for diagnosis only
UNIT READINESS	. Determined locally and based on evaluator judgment	. Same as ARTEP	. Evaluation based on analysis of objective data . Determined locally and based on evaluator judgment	. Determined locally and based on evaluator judgment related to a list of objectives	. Determined by cost and chievement trade-offs . No evaluator judgment included	. Same as UPAM	. Determined by unit mission performance . No evaluator judgment considered	. Same as TO but judgment on performance provided for diagnosis
SUBUNIT	. Not directly provided as part of Bn or Co exercise	. Same as ARTEP	. Permits determination of subunit readiness (for each participating echelon in context of overall exercise)	. Permits determination of subunit readiness (for each participating echelon in context of overall exercise)	. Not directly provided as part of Bn or Co exercise	. Same as UPAM	. Not directly provided as part of Bn or Co exercise	. Same as TO
DIAGNOSIS	. Only to the degree the evaluator selects . Inconsistent across evaluators Army-wide	. Same as ARTEP	. Training deficiencies identified consistently . Deficiencies determined by expert judgment also provided	. Training deficiencies related to specific objectives	. Not directly provided	. Same as UPAM	. Not directly provided	. Training deficiencies identified consistently . Deficiencies determined by expert judgment also provided

Table 1. Alternative ARTEP Models

ARTEP EVALUATION MODELS:	SOURCES OF DATA (INPUTS)			METHOD	OUTPUTS	
	PROCESS	INTER-MEDIATE	PRODUCT		READINESS	DIAGNOSIS
1. ARMY TRAINING AND EVALUATION PROGRAM (ARTEP)	X			Ev=EJ	X	X
2. ARTEP + ENGAGEMENT SIMULATION (ARTEP + ES)	X	X	X	Ev=EJ	X	X
3. SITUATION DEPENDENT MODEL (SIT. DEP)	X	X	X	Ev= [(EP+IP+P)*EJ]	X	X
4. OBJECTIVES BASED (OBJ. BAS)	X	X	X	Ev= [(EP=EPO) +(IP=IPO) ±(P=PO)]	X	X
5. UNIT PERFORMANCE ASSESSMENT MODEL (UPAM)			X	Ev=EP(UPAM)	X	
6. UPAM/F (UPAM/FIXED COST AND ACHIEVEMENT[1])			X	Ev=EP(UPAM)	X	
7. TERMINAL OBJECTIVES (TO)			X	Ev=(EP=EPO)	X	
8. TERMINAL OBJECTIVES +			X	Ev=(EP=EPO)+	X	
DIAGNOSIS (TOD)	X	X	X	D=EJ		X

[1] UPAM/F requirements fixed by higher level, i.e., DA, Schools, TRADOC.

SYMBOLS. P=PROCESS, IP=INTERMEDIATE PRODUCT, EP=END PRODUCT, Ev=EVALUATION, EJ=EXPERT JUDGEMENT, O=OBJECTIVE, D=DIAGNOSIS.

Figure A. Evaluation Models Presented as Functions of Inputs and Outputs.

Database	ARTEP	ARTEP/ES	SIT.DEP.	OBJ.BAS	UPAM	UPAM/F	T.O.	T.O.D.
(1) Subunit data used in evaluation: +/-								
Subunit data available from: +/-								
(2) Process data used in evaluation: +/-								
Process data available from: +/-								
(3) Cause/effect info. used in eval.: +/-								
Cause/effect info. available from: +/-								
(4) Cost vs. achvmt. info. used to eval.: +/-								
Cost vs. achvmt. info. available: +/-								
(5) Casualty asses. system: +/-								
Validity								
(6) Objectivity of scoring strategy: Hi/Med/Lo								
(7) Objectivity of product data: Hi/Med/Lo								
(8) Standards: Flexible/fix								
(9) How to perform info. provided: +/-								
(10) Clarity of passing requirement: Hi/Med/Lo								
(11) Gamesmanship potential: Hi/Med/Lo								
(12) Ease of interpretation: Hi/Med/Lo								
Implementability								
(13) # Controllers:								
(14) # Evaluators:								
(15) Evaluator training: Hi/Med/Lo								
(16) Controlled aggressor environ.: +/-								
(17) Flexibility (terr., weather, equip., pers.): Hi/Med/Lo								
(18) Ease of erployment: Hi/Med/Lo								
(19) Accep. to mil. trnr.: Hi/Med/Lo								
(20) Accep. to mil. eval.: Hi/Med/Lo								
Availability								
(21) Now/Soon/OTII-1/ OTII/OTII+								

Figure B. Criterial Framework Developed for Comparing ARTEP Evaluation Models

Database	ARTEP	ARTEP/ES	SIT.DEP.	OBJ.BAS	UPAM	UPAM/F	T.O.	T.O.D.
(1) Subunit data used in evaluation: +/-								
Subunit data available from: +/-								
(2) Process data used in evaluation: +/-								
Process data available from: +/-								
(3) Cause/effect info. used in eval.: +/-	X	X						
Cause/effect info. available from: +/-	X	X						
(4) Cost vs. achvmt. info. used to eval.: +/-								
Cost vs. achvmt. info. available: +/-								
(5) Casualty asses. system: +/-		X				X	X	X
Validity								
(6) Objectivity of scoring strategy: Hi/Med/Lo								
(7) Objectivity of product data: Hi/Med/Lo			X	X				
(8) Standards: Flexible/fix								
(9) How to perform info. provided: +/-								
(10) Clarity of passing requirement: Hi/Med/Lo				X	X	X		
(11) Gamesmanship potential: Hi/Med/Lo				X	X	X	X	X
(12) Ease of interpretation: Hi/Med/Lo								
Implementability								
(13) # Controllers:								
(14) # Evaluators:				X				X
(15) Evaluator training: Hi/Med/Lo	X	X		X				X
(16) Controlled aggressor environ.: +/-								
(17) Flexibility (terr., weather, equip., pers.): Hi/Med/Lo				X	X		X	X
(18) Ease of employment: Hi/Med/Lo	X							X
(19) Accep. to mil. trng.: Hi/Med/Lo	X	X		X				X
(20) Accep. to mil. eval.: Hi/Med/Lo		X		X				
Availability								
(21) Now/Soon/OTII-1/ OTII/OTII+								

*(X) indicates disagreement

Figure C. Disagreement among Ratings of 8 Models by Research Team

Database	ARTEP	ARTEP/ES	SIT.DEP.	OBJ.BAS	UPAM	UPAM/F	T.O.	T.O.D.
(1) Subunit data used in evaluation: +/-	-	-	+	+	-	-	-	-
Subunit data available from: +/-	-	-	+	+	-	-	-	+
(2) Process data used in evaluation: +/-	+	+	+	+	-	-	-	-
Process data available from: +/-	+	+	+	+	-	-	-	+
(3) Cause/effect info. used in eval.: +/-	-	-	+	-	-	-	-	-
Cause/effect info. available from: +/-	-	-	+	+	-	-	-	+
(4) Cost vs. achvmt. info. used to eval.: +/-	-	-	+	-	+	+	-	-
Cost vs. achvmt. info. available: +/-	-	-	+	-	+	+	-	-
(5) Casualty asses. system: +/-	-	+	+	+	+	+	+	+
Validity								
(6) Objectivity of scoring strategy: Hi/Med/Lo	L	L	M	H	H	H	H	H
(7) Objectivity of product data: Hi/Med/Lo	L	M	H	H	H	H	H	H
(8) Standards: Flexible/fix	Fl	Fl	Fl	Fx	Fl	Fx	Fx	Fx
(9) How to perform info. provided: +/-	+	+	+	+	-	-	-	-
(10) Clarity of passing requirement: Hi/Med/Lo	L	L	M	H	M	M	H	H
(11) Gamesmanship potential: Hi/Med/Lo	H	M	L	M	H	H	H	H
(12) Ease of interpretation: Hi/Med/Lo	L	M	M	H	H	H	H	H
Implementability								
(13) # Controllers:	1	15	15	15	22	22	22	22
(14) # Evaluators:	3	8	8	8	-	-	-	-
(15) Evaluator training: Hi/Med/Lo	M	M	H	M	L	L	L	H
(16) Controlled aggressor environ.: +/-	+	+	-	+	-	+	+	+
(17) Flexibility (terr., weather, equip., pers.): Hi/Med/Lo	M	M	H	L	M	L	L	L
(18) Ease of employment: Hi/Med/Lo	H	M	M	M	H	H	H	M
(19) Accep. to mil. trng.: Hi/Med/Lo	L	M	H	M	L	L	L	H
(20) Accep. to mil. eval.: Hi/Med/Lo	L	M	M	M	H	H	H	H
Availability								
(21) Now/Soon/OTII-1/ OTII/OTII+	Now	Soon	OTII	OTII+	OTII-1	OTII-1	OTII-1	OTII

Figure D. Final Ratings of 8 Models by Research Team

Database	ARTEP	ARTEP/ES	SIT.DEP.	OBJ.BAS	UPAM	UPAM/F	T.O.	T.O.D.
(1) Subunit data used in evaluation: +/-								
Subunit data available from: +/-								X
(2) Process data used in evaluation: +/-								
Process data available from: +/-								
(3) Cause/effect info. used in eval.: +/-		X*		X				X
Cause/effect info. available from: +/-		X						
(4) Cost vs. achvmt. info. used to eval.: +/-		X		X			X	X
Cost vs. achvmt. info. available: +/-		X		X			X	X
(5) Casualty asses. system: +/-								
Validity								
(6) Objectivity of scoring strategy: Hi/Med/Lo								
(7) Objectivity of product data: Hi/Med/Lo								
(8) Standards: Flexible/fix								
(9) How to perform info. provided: +/-								X
(10) Clarity of passing requirement: Hi/Med/Lo								
(11) Gamesmanship potential: Hi/Med/Lo					X	X	X	X
(12) Ease of interpretation: Hi/Med/Lo					X	X	X	
Implementability								
(13) # Controllers:								
(14) # Evaluators:								
(15) Evaluator training: Hi/Med/Lo								
(16) Controlled aggressor environ.: +/-			X		X			
(17) Flexibility (terr., weather, equip., pers.): Hi/Med/Lo				X				
(18) Ease of employment: Hi/Med/Lo					X	X		
(19) Accep. to mil. trnr.: Hi/Med/Lo								
(20) Accep. to mil. eval.: Hi/Med/Lo					X	X	X	
Availability								
(21) Now/Soon/OTII-1/ OTII/OTII+								

*(X) indicates disagreement

Figure E_1. Variance between Decision-Maker$_1$ and Research Team on Ratings of 8 Models

Database	ARTEP	ARTEP/ES	SIT.DEP.	OBJ.BAS	UPAM	UPAM/F	T.O.	T.O.D.
(1) Subunit data used in evaluation: +/-								
Subunit data available from: +/-								
(2) Process data used in evaluation: +/-								
Process data available from: +/-								
(3) Cause/effect info. used in eval.: +/-		X*	X					
Cause/effect info. available from: +/-								
(4) Cost vs. achvmt. info. used to eval.: +/-								
Cost vs. achvmt. info. available: +/-								
(5) Casualty asses. system: +/-								
Validity								
(6) Objectivity of scoring strategy: Hi/Med/Lo								
(7) Objectivity of product data: Hi/Med/Lo								
(8) Standards: Flexible/fix								
(9) How to perform info. provided: +/-								
(10) Clarity of passing requirement: Hi/Med/Lo								
(11) Gamesmanship potential: Hi/Med/Lo								
(12) Ease of interpretation: Hi/Med/Lo								
Implementability								
(13) # Controllers:								
(14) # Evaluators:								
(15) Evaluator training: Hi/Med/Lo								
(16) Controlled aggressor environ.: +/-								
(17) Flexibility (terr., weather, equip., pers.): Hi/Med/Lo								
(18) Ease of employment: Hi/Med/Lo								
(19) Accep. to mil. trng.: Hi/Med/Lo								X
(20) Accep. to mil. eval.: Hi/Med/Lo								
Availability								
(21) Now/Soon/OTII-1/ OTII/OTII+								

*(X) indicates disagreement

Figure E_2. Variance between Decision-Maker$_2$ and Research Team on Ratings of 8 Models

Databases	ARTEP	ARTEP/ES	SIT.DEP.	OBJ.BAS	UPAM	UPAM/F	T.O.	T.O.D.
(1) Subunit data used in evaluation: +/-								
Subunit data available from: +/-								
(2) Process data used in evaluation: +/-								
Process data available from: +/-								
(3) Cause/effect info. used in eval.: +/-			X*	X	X	X		X
Cause/effect info. available from: +/-			X					
(4) Cost vs. achvmt. info. used to eval.: +/-			X					
Cost vs. achvmt. info. available: +/-			X					
(5) Casualty asses. system: +/-								
Validity								
(6) Objectivity of scoring strategy: Hi/Med/Lo								
(7) Objectivity of product data: Hi/Med/Lo				X				
(8) Standards: Flexible/fix								X
(9) How to perform info. provided: +/-								
(10) Clarity of passing requirement: Hi/Med/Lo								
(11) Gamesmanship potential: Hi/Med/Lo			X		X	X		
(12) Ease of interpretation: Hi/Med/Lo								
Implementability								
(13) # Controllers:								
(14) # Evaluators:								
(15) Evaluator training: Hi/Med/Lo								X
(16) Controlled aggressor environ.: +/-			X		X			
(17) Flexibility (terr., weather, equip., pers.): Hi/Med/Lo								
(18) Ease of employment: Hi/Med/Lo					X	X		
(19) Accep. to mil. trng.: Hi/Med/Lo								X
(20) Accep. to mil. eval.: Hi/Med/Lo								
Availability								
(21) Now/Soon/OTII-1/ OTII/OTII+								

* (X) indicates disagreement

Figure E_3. Variance between Decision-Maker$_3$ and Research Team on Ratings of 8 Models

Database	ARTEP	ARTEP/ES	SIT.DEP.	OBJ.BAS	UPAM	UPAM/F	T.O.	T.O.D.
(1) Subunit data used in evaluation: +/-	*							
Subunit data available from: +/-								
(2) Process data used in evaluation: +/-								
Process data available from: +/-								
(3) Cause/effect info. used in eval.: +/-								
Cause/effect info. available from: +/-								
(4) Cost vs. achvmt. info. used to eval.: +/-								
Cost vs. achvmt. info. available: +/-								
(5) Casualty asses. system: +/-								
Validity								
(6) Objectivity of scoring strategy: Hi/Med/Lo								
(7) Objectivity of product data: Hi/Med/Lo								
(8) Standards: Flexible/fix								
(9) How to perform info. provided: +/-								
(10) Clarity of passing requirement: Hi/Med/Lo								
(11) Gamesmanship potential: Hi/Med/Lo								
(12) Ease of interpretation: Hi/Med/Lo								
Implementability								
(13) # Controllers:								
(14) # Evaluators:								
(15) Evaluator training: Hi/Med/Lo								
(16) Controlled aggressor environ.: +/-								
(17) Flexibility (terr., weather, equip., pers.): Hi/Med/Lo								
(18) Ease of employment: Hi/Med/Lo								
(19) Accep. to mil. trng.: Hi/Med/Lo								
(20) Accep. to mil. eval.: Hi/Med/Lo								
Availability								
(21) Now/Soon/OTII-1/ OTII/OTII+								

* (No disagreements)

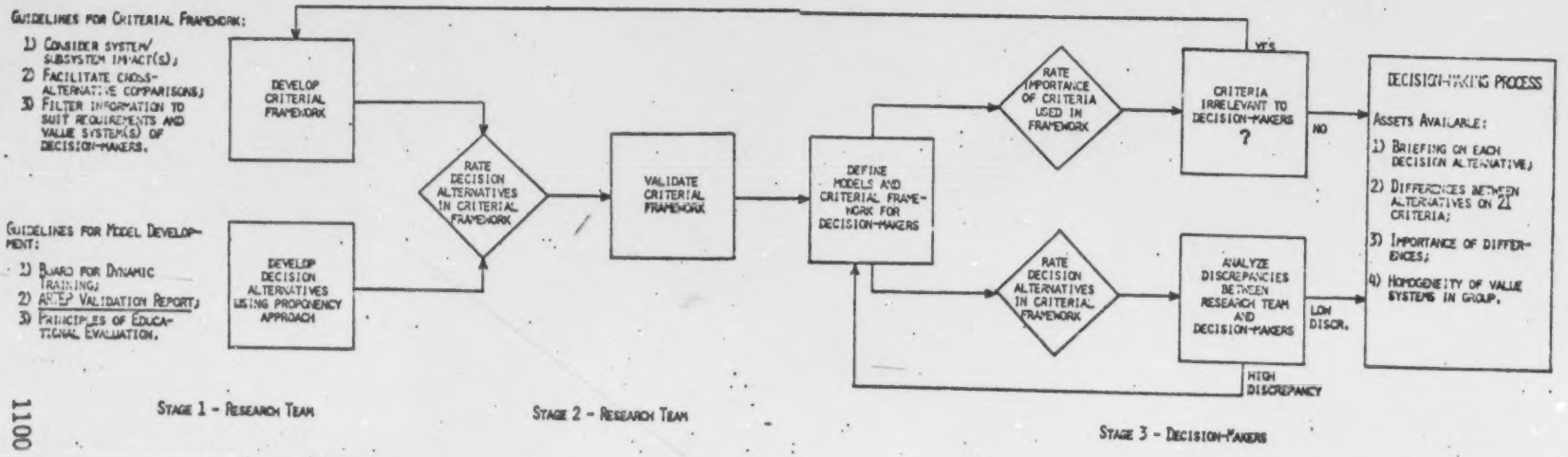
Figure E_4. Variance between Decision-Maker$_4$ and Research Team on Ratings of 8 Models

	D_1	D_2	D_3	D_4			X
Database							
(1) Submit data used in evaluation: +/-	3	1	4	3			2.7
Submit data available from: +/-	3	1	4	4			3.0
(2) Process data used in evaluation: +/-	4	4	1	0			2.2
Process data available from: +/-	4	4	1	1			2.5
(3) Cause/effect info. used in eval.: +/-	5	4	5	4			4.5
Cause/effect info. available from: +/-	5	4	5	4			4.5
(4) Cost vs. achvmt. info. used to eval.: +/-	2	1	4	3			2.5
Cost vs. achvmt. info. available: +/-	2	1	4	4			2.7
(5) Casualty asses. system: +/-	4	2	3	3			3.0
Validity							
(6) Objectivity of scoring strategy: Hi/Med/Lo	3	3	4	4			3.5
(7) Objectivity of product data: Hi/Med/Lo	3	3	4	3			3.2
(8) Standards: Flexible/fix	3	4	3	3			3.2
(9) How to perform info. provided: +/-	5	4	4	4			4.3
(10) Clarity of passing requirement: Hi/Med/Lo	5	4	4	4			4.3
(11) Gamesmanship potential: Hi/Med/Lo	3	3	2	3			2.7
(12) Ease of interpretation: Hi/Med/Lo	5	2	5	5			4.3
Implementability							
(13) # Controllers:	3	3	3	3			3.0
(14) # Evaluators:	3	4	3	3			3.2
(15) Evaluator training: Hi/Med/Lo	5	3	4	4			4.0
(16) Controlled aggressor environ.: +/-	3	3	4	2			3.0
(17) Flexibility (terr., weather, equip., pers.): Hi/Med/Lo	4	4	2	3			3.2
(18) Ease of employment: Hi/Med/Lo	4	4	3	3			3.5
(19) Accep. to mil. trnr.: Hi/Med/Lo	5	4	5	5			4.8
(20) Accep. to mil. eval.: Hi/Med/Lo	4	3	5	5			4.3

(Item 21 not judged as to importance)

0=irrelevant; 1=not very important; 2=fairly important; 3=important; 4=very important; 5=critically important.

Figure F. Importance katings for 21 Criteria by Decision-Makers D_1 to D_4.


GUIDELINES FOR CRITERIAL FRAMEWORK:
1) CONSIDER SYSTEM/SUBSYSTEM IMPACT(S);
2) FACILITATE CROSS-ALTERNATIVE COMPARISONS;
3) FILTER INFORMATION TO SUIT REQUIREMENTS AND VALUE SYSTEM(S) OF DECISION-MAKERS.
GUIDELINES FOR MODEL DEVELOPMENT:
1) BOARD FOR DYNAMIC TRAINING;
2) ARTEP VALIDATION REPORT;
3) PRINCIPLES OF EDUCATIONAL EVALUATION.
DEVELOP CRITERIAL FRAMEWORK
DEVELOP DECISION ALTERNATIVES USING PROPONENCY APPROACH
RATE DECISION ALTERNATIVES IN CRITERIAL FRAMEWORK
VALIDATE CRITERIAL FRAMEWORK
DEFINE MODELS AND CRITERIAL FRAMEWORK FOR DECISION-MAKERS
RATE IMPORTANCE OF CRITERIA USED IN FRAMEWORK
RATE DECISION ALTERNATIVES IN CRITERIAL FRAMEWORK
CRITERIA IRRELEVANT TO DECISION-MAKERS ?
YES
NO
ANALYZE DISCREPANCIES BETWEEN RESEARCH TEAM AND DECISION-MAKERS
LOW DISCR.
HIGH DISCREPANCY
DECISION-MAKING PROCESS
ASSETS AVAILABLE:
1) BRIEFING ON EACH DECISION ALTERNATIVE;
2) DIFFERENCES BETWEEN ALTERNATIVES ON 21 CRITERIA;
3) IMPORTANCE OF DIFFERENCES;
4) HOMOGENEITY OF VALUE SYSTEMS IN GROUP.
STAGE 1 - RESEARCH TEAM
STAGE 2 - RESEARCH TEAM
STAGE 3 - DECISION-MAKERS

1100

Figure G. Schematic Showing Evaluation Methodology Followed in Aiding Decision-Making on ARTEP Model for MILES OT II

ACHIEVEMENT TESTING FOR VARIOUS TYPES OF CONCENTRATED MANAGEMENT COURSES

Prepared for the
18th Conference of the
Military Testing Association

by

John F. McAreavy, Ph. D.
Department of Public Administration
U.S. Army Management Engineering Training Activity

October 1976

1220

1101

ACHIEVEMENT TESTING FOR VARIOUS TYPES OF CONCENTRATED MANAGEMENT COURSES

Introduction

In order to satisfy organizational and individual development needs, government, business and industry provide employees with extensive opportunities for management and management specialist training. This training may take various forms depending on the specific educational objectives, availability of educational resources and urgency. One manner of satisfying these developmental needs is to design and conduct short concentrated courses of instruction in the management and management specialist areas. Generally, these courses meet six to eight hours a day over a three to fifteen day period. The costs associated with these courses are significant when one considers the resources expended on course development, presentation, facilities and the salaries of the attendees. It is therefore incumbent on managers of the attendees and the management of the training institutions to make every effort possible to maximize the relevance, efficiency and effectiveness of such courses.

Relevancy, Efficiency and Effectiveness

To assure the relevancy of a course of instruction, there must be a candid exchange of information between the operational personnel and the training personnel. Specific operational deficiencies and employee needs must be articulated and translated into goals that can be achieved by training. The output of this dialogue must be clear, concise training objectives stated in behavioral terms in the operational setting.

In order to maximize training efficiency, the training organization must be aware of all of the alternate modes of instruction and their advantages in achieving various kinds of educational objectives. With this information, the course material can be presented in the most appropriate instructional mode so that the attendees' time away from the job will be efficiently utilized.

Assuming that the course objectives are relevant to operational needs and efficient training modes are selected for course material presentation, it is of paramount

importance that the effectiveness of the course be measured in terms of individual performance and this information provided to the attendee's supervisor in order that the employee can be assigned tasks that he is capable of performing. This paper is directed toward the identification of the most appropriate methods of evaluating the effectiveness of student performance in various types of concentrated management and management analysis courses.

Types of Courses

After thorough consideration of the relevance and efficiency factors, there appears to be three general types of formal classroom instruction in the management and management analysis areas: orientations, technical courses and seminars.

Orientation courses are designed to satisfy operational needs that are geared to improve coordination, cooperation and communications within and among organizations. They are characterized by introductory level information about a subject to (1) familiarize new employees with the subject matter area, (2) to retrain personnel who were formally working in the subject matter area, or (3) to provide "nice to know" or enrichment information to employees.

Technical courses are designed to satisfy operational needs involving the acquisition of knowledge, skills and abilities of analysts, designers, auditors and planners in order that they might accomplish their assigned tasks in a proficient manner. Emphasis is on evaluation, analysis, synthesis and implementation.

Seminars are designed to provide a forum for employees in similar positions to share ideas regarding operational problems involved in increasing organizational productivity and improving individual job performance. The principle emphasis in the seminars is to generate professional or technical cohesiveness that will spawn innovation and cooperation in achieving the organization's goals.

Classroom Environment

The orientation classroom is characterized by the dissemination of basic knowledge and applications through lecture discussion and various audio visual presentations. Comprehension is enhanced by student participation in individual and group

exercises illustrating the application of knowledge in diversified operational settings. The technical course classroom is characterized by a three-phase process of (1) knowledge dissemination to the student, (2) a presentation of an application of the knowledge by the instructor, and (3) the assignment of an individual or group problem for student solution. The instructor presents the material from a background of experience in the technical or professional field. The classroom environment in the seminars are characterized by case studies, role playing, simulation, panels and discussion groups. The instructor takes on the role of principally a moderator/facilitator rather than a presenter/instructor.

Since these three types of courses vary to such a wide degree in purpose, type of student and classroom environment, it should be obvious that the evaluation of the effectiveness of these three types of courses would also vary. The subsequent discussion will highlight the evaluation process applicable to technical courses and seminars but concentrate on the process as it applies to orientation courses.

Developing Evaluation Instruments

The fundamental goal of the designer of an evaluation instrument is to make the elements of the instrument as nearly equivalent to the elements of the operational setting as considerations of efficiency, practicability and expediency will permit. With this in mind, it is apparent that evaluation instruments will differ depending upon the type of course. It is also apparent that there would be a certain variety of instruments utilized within each course type, but the variety among course types would be more pronounced.

Regardless of the type of course, the evaluation of student performance involves three sub-processes:

- Developing evaluation instruments that can best measure the achievement of course objectives.
- Assessing the validity and reliability of the evaluation instruments.
- Interpreting the evaluation results in operational terms for the employee and the supervisor.

Evaluation by Type of Course

The evaluation instruments most appropriate for the technical courses are those that possess as many characteristics as possible to those of the operational setting. Since these courses are designed to provide the attendee with the skill and knowledge to accomplish specific tasks at their job, the evaluation instruments should consist of as many "identical elements" as possible. In the case of a COBOL computer programming course, the evaluation would consist of exercises presented to the student that would be nearly identical to the tasks assigned to the employee by his supervisor. This being the case, it can be seen just how important it is for the evaluation designer to know exactly what the supervisor expects from the student upon his return from the course. The more precise information that the supervisor can provide with respect to on-the-job task performance, the more precise information the supervisor can receive with respect to the employee's expected performance on the job upon the completion of the course. In general, the evaluation instruments most appropriate for this type of course are "practical" or "laboratory" examinations.

The seminar type of course is designed to achieve broader, less precise objectives than those specified for technical courses. Seminars are geared to attain objectives in the analysis, synthesis and evaluation subsets of the cognitive domain[1] and various subsets of the affective domain.[2] Evaluation instruments utilized in these areas are characterized by direct instructor observation of simulated operational situations. Role-playing exercises, incident processes and in-basket routines have been effective in the management and management specialist areas. Paper and pencil tests can also be effective if they are carefully constructed. Short case studies requiring the student to analyze a management situation and generate a solution involving synthesis and

---

[1] Bloom, Benjamin S. (ed.). Taxonomy of Educational Objectives. New York: Longmans, Green & Co., 1956.

[2] Krathwohl, D. R., et al. Taxonomy of Educational Objectives: Affective Domain. New York: David McKay Company, Inc., 1964.

and evaluation abilities have been quite successful. These tests have the added advantage of being more efficient when class sizes are large.

Orientation courses are designed to provide the attendee with knowledge about management philosophies, concepts, systems, methods and techniques. In addition, course objectives involving the comprehension and application[3] subsets of the cognitive domain are specified for attainment during the course of instruction. This type of course is especially suited for the more traditional forms of achievement tests consisting of multiple choice, matching and completion items.

Operational Needs

The first event in the overall process occurs when operations managers perceive deficiencies that inhibit the attainment of the organization's goals and objectives. For example:

- Production is decreasing because of bottlenecks caused by inadequate systems and procedures.
- Cooperation and coordination is weak between the Production Control Department and the Computer Department caused by lack of understanding of computer operations by the Production Control personnel, and vice versa.
- Lack of initiative and inadequate knowledge of contemporary industry practice in the sales department is causing a decline in the company's share of the market.

The articulation of these deficiences provide the basis for establishing operational needs and, in turn, course objectives for specific courses.

Course Objectives

Close coordination between operational personnel and training personnel will yield general course objectives. At this stage, these objectives can be formulated as a course description that includes as a minimum: Purpose, Scope and Student

---

[3] Bloom, op. cit.

Prerequisites. This course description must serve as the linkage between the operations and training personnel. It will not completely specify operational needs nor instructional objectives, but will serve as a common bond between the two groups.

A description of an orientation course in Automatic Data Processing might take the following form:

AUTOMATIC DATA PROCESSING ORIENTATION

LENGTH: 1 Week

PURPOSE: This course is designed to provide managers, functional specialists, and other users of computers with the principles of computer data processing and how the management process can be facilitated through the use of computer-based information systems. Upon completion of this training, the attendee will know the functions of a computer as it processes business data to produce information for improved management. The student will understand the participative role of the user during the development of the computer-based information system and the basic considerations during systems design.

SCOPE: Computer principles will be presented so that specific functions of input, storage, control, and output are understood as it occurs during computer data processing. Management Information Systems (MIS) concepts and goals will be explained as they relate to the mission of the organization including consideration of the data base design to support the systems information objectives. Details of systems analysis techniques and computer program development will be presented to the degree necessary for the user of computer products to understand what must be done to satisfy an information requirement. The activities of the computer center will be discussed to identify the functions it performs in processing data for the user and in controlling the data base for the manager. Topics pertaining to computer systems management will include a listing of steps, work processes, relationships, and considerations that must be made during the accomplishment of a computer automation project.

PREREQUISITES: This course is designed for functional specialists and middle and top managers having the rank of Captain (0-3) or civilian (GS-9) and above. Trainees

who require this course as a prerequisite for further training or career development but do not meet the grade requirement may be permitted to attend. Candidates should be from organizations which are potential or current users of ADP and have a need for this training in their work assignments. No knowledge of electronics or any special preparation in mathematics is necessary.

Instructional Objectives

The course description forms the basis for the specification of instructional objectives. When developing a list of general instructional objectives, the aim is to obtain a list of goals to work toward and not a list of specific types of behavior. Obviously, each general instructional objective will need to be defined further by a sample of the specific types of behavior that characterize each objective, but at this stage, the focus is on the general goal-directed objectives. Since the statement of instructional objectives is of primary importance, it is necessary that one understand what constitutes a valid objective and what constitutes a faulty objective. Some faults typically found in statements of course objectives are as follows:

The objective is one that cannot be achieved and certainly cannot be evaluated within a unit, i.e.,

DEVELOPS AN ABIDING INTEREST IN MANAGEMENT ENGINEERING TECHNIQUES THROUGHOUT HIS CAREER.

The objective is expressed in terms of unobservables, i.e.,

DEVELOPS AN APPRECIATION OF AUTOMATIC DATA PROCESSING.

The objective bears little or no relationship to the content of instruction, i.e.,

CAN PROVE THE CENTRAL LIMIT THEORM.

The statement of the objective implies a process quite different from the one that is taught, i.e.,

APPRECIATES THE THEORY OF RELATIVITY.

The objective may be inappropriate to the level of instruction, i.e.,

CAN SOLVE SECOND ORDER, THIRD DEGREE DIFFERENTIAL EQUATIONS.

In order to specify meaningful instructional objectives, it is advantageous to separate the instructional process in the course from the content of the course.

Taxonomies of educational objectives in both the cognitive[4] and affective[5] domains have been developed that provide tractable models for establishing instructional objectives. The categories for classifying objectives in each domain are arranged in hierarchical order, from the simplest behavior outcomes to the most complex. For example, the cognitive domain starts with simple knowledge outcomes and then proceeds through increasingly complex levels. Each category is assumed to include the behavior at the lower levels. The affective domain follows a similar pattern. The categories, a brief description of each, and selected subcategories of the taxonomy in the cognitive domain follows:

TAXONOMY OF
EDUCATIONAL OBJECTIVES
(Cognitive Domain)

Code	Category-Subcategory	Brief Description
1.00	Knowledge	The remembering of previously learned material.
	1.10 Knowledge of Specifics	
	1.20 Knowledge of Ways and Means of Dealing with Specifics	
	1.30 Knowledge of the Universals and Abstractions in a Field	
2.00	Comprehension	The ability to grasp the meaning of the material.
	2.10 Translation	
	2.20 Interpretation	
	2.30 Extrapolation	
3.00	Application	The ability to use learned material in new and concrete situations.
4.00	Analysis	The ability to break down material into its component parts so that its organizational structure may be understood.

[4] Ibid.

[5] Krathwohl, op. cit.

Code	Category-Subcategory	Brief Description
5.00	Synthesis	The ability to put parts together to form a new whole.
6.00	Evaluation	The ability to judge the value of material based on definite criteria.

By utilizing the listing of operational needs, the course description and the information in the Taxonomy of Educational Objectives for the cognitive domain, the objectives for a short orientation course in ADP might be stated as follows:

1. Know the fundamental facts, concepts, principles, methods and techniques of ADP.

2. Know the fundamental purposes and interrelationships of computer functions, capabilities, and applications.

3. Understands the role of the computer in data processing and management.

In addition to these objectives, the subject matter to be covered in the course must be defined. In the case of the ADP orientation course, the content was grouped into five basic subject matter areas:

History of Computers and Management
Computer Functions
Computer Programming
Systems Analysis
Management Applications

Evaluation (Test) Objectives

The process objectives merged with the subject matter content forms the basis for the course development and the specifications of test objectives. Using the ADP course example, a listing of test objectives would include, but not be limited to:

AUTOMATIC DATA PROCESSING ORIENTATION
TEST OBJECTIVES (A Selection)

1.11 Knowledge of Terminology
. Define computer and data processing terms by giving their attributes, properties, or relationships.

1.12 Knowledge of Specific Facts
. Recall or recognize factual information about computers.

1.21 Knowledge of Conventions

. Recognize correct forms, methods, techniques used in computer technology.

1.22 Knowledge of Trends and Sequences

. Specify the proper sequence of activities in ADP applications.

1.23 Knowledge of Classifications and Categories

. Recognize the features and functions of the various forms of ADP operations and equipment.

1.24 Knowledge of Criteria

. Specify proper criteria for judging the effectiveness and/or efficiency of an existing or planned ADP application.

1.25 Knowledge of Methodology

. Recognize the techniques and methods used by ADP personnel in determining feasibility, practicability, analysis, design, and implementing ADP applications.

1.31 Knowledge of Principles and Generalizations

. Recall fundamental principles and generalizations involved in ADP.

1.32 Knowledge of Theories and Structures

. Recognize fundamental theories and structures in the ADP field.

2.10 Translation

. Demonstrate the ability to translate a problem given in technical phrasecology into concrete or less abstract phrasecology.

. Demonstrate the ability to translate an abstraction, such as a general principle, by giving an illustration or sample.

2.20 Interpretation

. Demonstrate the ability to distinguish among warranted, unwarranted, or contradicted conclusions drawn from a set of ADP and management data.

2.30 Extrapolation

. Demonstrate the ability to deal with conclusions of an ADP study in terms of inferences made from explicit statements.

3.00 Application

. Demonstrate the ability to apply principles, methods, and techniques to new situations.

Evaluation Blueprint

The development of an evaluation blueprint integrates the operational, course, instructional and evaluation objectives and provides a relevant basis for designing

an evaluation instrument that can effectively measure the individual's achievement of these objectives.

The blueprint is constructed in matrix form with the categories of subject matter on the horizontal axis and the categories of the taxonomy of educational objectives on the vertical axis. For the orientation course in ADP, the horizontal axis would consist of five columns: History of Computers and Management, Computer Functions, Computer Programming, Systems Analysis and Management Applications, and the vertical axis would consist of three rows: Knowledge, Comprehension and Application.

Upon examining the specific objectives and the amount of time devoted to each subject matter area, the marginal percentages can be placed on the matrix. In the ADP orientation course, 70% of the objectives fall in the Knowledge category, 26% fall in the Comprehension category and 4% fall in the Application category. Also, the percentage of time expended in the five subject matter areas was determined to be: History of Computers and Management, 8%; Functions of Computers, 32%; Computer Programming, 20%; Systems Analysis, 20%; Management Applications, 20%. The elements in the matrix are computed by multiplying the row percentage by the corresponding column percentage, i.e., the Comprehension/System Analysis element is 5.2% (the product of 26% and 20%). This element percentage, then, represents the percentage of evaluation effort that should be devoted to the objectives relating to the row and column headings of the element. In the case of a multiple choice examination, 5% of the test items would be directed toward the comprehension of systems analysis. An evaluation blueprint for an ADP orientation course is presented in the following table.

The blueprint for a technical course would undoubtedly include the same categories of educational objectives, i.e., Knowledge, Comprehension and Application, but would yield larger percentages in the Comprehension and Application categories and less in the Knowledge category. The blueprint for a seminar course would also include the Analysis, Synthesis and Evaluation categories of the Taxonomy of Educational Objectives. The percentages in each category for the seminar and technical course blueprints would

AN EVALUATION BLUEPRINT

for an

AUTOMATIC DATA PROCESSING ORIENTATION COURSE

Percentage of Evaluation Effort

Educational Objectives \ Subject Matter Area	History of Computers and Management	Functions of Computers	Programming	Systems Analysis	Management Applications	Totals
Knowledge	5.6%	22.4%	14.0%	14.0%	14.0%	70%
Comprehension	2.1%	8.3%	5.2%	5.2%	5.2%	26%
Application	.3%	1.3%	.8%	.8%	.8%	4%
Totals	8%	32%	20%	20%	20%	100%

1113

1232

1233

be determined by analyzing the specific objectives for each course, as was done in developing the blueprint for the ADP orientation course.

Construction of the Evaluation Instrument

The specific procedures for constructing the components of the evaluation instruments are available from a number of sources[6], therefore this topic will not be addressed in this paper. It must be emphasized, however, that these procedures should be applied in such a fashion that the evaluation instrument conforms to the blueprint generated for each course.

Validity and Reliability of the Evaluation Instrument

The student's performance in the course will be determined in large measure by the performance in the prepared evaluation instruments. Course performance will be provided to the operational supervisor who will rely on this information to assign appropriate tasks to the employee. It is therefore imperative that the instruments are valid and reliable.

Validity refers to the extent to which a test actually measures what it is intended to measure. The key to validity is the proper construction of the test blueprint. By properly using the procedures discussed in this paper, one is assured that the test possesses the required validity.

Reliability pertains to the degree to which a test can yield consistent results under similar conditions. Reliability can best be measured by utilizing equivalent test forms and can best be improved by analyzing three factors: item difficulty, number of items and testing conditions.[7]

Interpreting Results to Students and Supervisors

The results of an evaluation should be discussed in detail with the students. In most cases, this can be accomplished by reviewing each item in a group session. The most effective way to transmit information about student performance to the supervisor is by a brief essay appraisal. This approach, however, is not practical

---

[6] Lindquist, E. F. (ed.). Educational Measurement. Washington, DC: American Council on Education, 1951.

[7] Ibid, Chapter 15.

in many cases, so an efficient alternative is to relate student test performance to a grading system that briefly explains the meaning of the grade to the supervisor. An abbreviated example of this approach could take the form:

Grade	Explanation
A	Demonstrated outstanding comprehension of course principles and techniques and an exceptional facility to apply them.
B	Demonstrated above average comprehension of course principles and techniques and a commendable facility to apply them.
C	Demonstrated average comprehension of course principles and techniques and an apparent facility to apply them.
D	Demonstrated a limited comprehension of course principles and techniques. Opportunities for application should be carefully selected and supervision provided.
F	Demonstrated a lack of comprehension of the subject matter presented to the degree necessary for its use.
I	Student was unable to complete the course due to absence from class.

This approach, although not as effective as the essay appraisal, is a vast improvement over systems that provide percentage grades, numerical grades or unexplained letter grades, which have little or no operational meaning to the supervisor.

Summary

As was indicated at the outset, in-service training is a significant expense in most organizations and, in order for their expense to pay dividends, the training must be relevant, efficient and effective. Effectiveness of management and management specialist training can be optimized only if operational needs can be articulated and translated into course, instructional and evaluation objectives, each in turn. The evaluation blueprint, reflecting these objectives, forms the basis for the construction of evaluation instruments that can provide meaningful information concerning the predicted performance of the employee in the operational setting.

References

Bloom, Benjamin S. (ed.). Taxonomy of Educational Objectives. New York: Longmans, Green & Co., 1956.

Gage, Nathaniel L. (ed.). Handbook of Research on Teaching. Chicago: Rand McNally & Co., 1963.

Lindquist, E. F. (ed.). Educational Measurement. Washington, DC: American Council on Education, 1951.

McKinney, J. F. An Analysis of Factors Affecting the Achievement of [illegible] Delegate in Short Concentrated Courses. The University of [illegible], Iowa City, Iowa. Unpublished Dissertation, 1969.

Thorndike, R. L., and Hagen, E. Measurement and Evaluation in Psychology and Education. New York: John Wiley & Sons, Inc., 1961.

Travers, Robert M. W. An Introduction to Educational Research. New York: Macmillan Co., 1966.

The Evaluation-in-Systems (EIS) Approach in Development of Training Programs

Maurice Dutton, Ph.D.
ARBEC, Inc.

and

Paul Liberty, Ph.D.
University of Texas
at Austin and
ARBEC, Inc.

Paper presented at the 18th Annual Conference of the Military Testing Association, Gulf Shores, Alabama, October 21, 1976.

INTRODUCTION

The military establishment has historically been on the forefront in the development of training programs and in understanding the relationship between good training and the achievement of military success. What has changed over the years is the increased complexity involved in identifying and achieving military objectives and in the cost and sophistication of utilizing complex weapons systems and counter-systems. This has led, in turn, to increased complexity in identifying training objectives and in the cost and sophistication of designing, developing, implementing, and conducting training programs.

This problem has forced program designers and developers to adopt increasingly sophisticated management tools and techniques. These techniques provide greater specificity in problem identification, objective setting, resource allocation, and evaluation techniques that improve monitoring and that identify and correct problems in their early stages. Accountability demands, as well as the complexity of military systems dictate, that we (society) can no longer afford the costs associated with ineffective and/or inefficient training design and development. We think the Evaluation-In-Systems approach used by ARBEC, Inc., helps meet the challenge for developing modern military training programs.

This paper describes a comprehensive systems approach to the development of training programs which incorporates systems

management techniques and strongly emphasizes the role of evaluation throughout the entire process. This systems approach was utilized over a three-year development period, including pilot and field testing, of four different training programs for use by counselors, supervisors, co-workers, and families associated with handicapped workers. In addition, a relatively unique program development team will be described which consisted of a consortium of grantor staff, contractor staff, and external evaluator.

The paper specifically traces the developmental process through the evaluation over seven (7) levels or types of evaluation: needs specification, context assessment, strategy selection, program planning, implementation evaluation, process evaluation, and product evaluation. Training manuals are available for examination and study by conference participants, along with the results of criterion-referenced tests and attitude measures used in the programs to promote affirmative action and facilitate selective placement of handicapped persons.

HISTORICAL SCOPE

In November of 1973, the Texas Rehabilitation Commission (TRC) issued a Request for Proposals for the development of a training program to train governmental supervisors in the use of qualified handicapped persons as employees. ARBEC, Inc., an Austin, Texas consulting firm, submitted a proposal and was subsequently awarded the contract. This contract was unusual in that a task force of TRC staff (governmental agency) and ARBEC staff (private, profit firm) was formed and actually worked as a

team throughout the entire contact period. The training program (product) was not a "turn key" job done by an "outside firm" but was a collaborative product based on the best that both a governmental agency and a private business could bring together.

The primary curriculum designers from both TRC and ARBEC had strong systems management backgrounds and found that to be a natural vehicle for their communication and approach to tasks. As a result, contracts and agreements focused on products and sub-products to be produced according to predetermined specifications and time lines. There was also a common desire to bring as much specificity as possible to such items as program goals and learning unit objectives, learning participant population identification, describing the typical setting in which learning would take place, identification of typical instructor qualifications, and determining the amount of released time possible for training. It was felt that if sufficient data could be collected on the above, then it would be possible to identify the most appropriate teaching/learning strategies and techniques and to develop the instructional materials needed. It was also agreed that evaluative data would be identified and collected for every step in process so that proper and timely decision-making could occur and problems could be identified and corrected quickly.

PROCEDURAL APPROACH

The following is an outline of the Evaluation-In-Systems approach to training program development and implementation that

evolved and was refined over a three-year, four-project period.

I. Research Phase

A. Problem Identification (Needs Specification)

1. Statement of felt problem

a. Population affected

b. Need not being met (or inadequately met)

c. Setting or situation in which problem occurs

2. Needs assessment to document problem

a. Direct data sources (surveys, etc.)

b. Indirect sources (review of literature, etc.)

B. Resource Identification (Context Assessment)

1. Identification of resources presently being expended on identified problem

2. Identification of resources immediately available which can be expended and other possible resources.

3. Specification of the socio-politic factors in the development context

C. Possible Problem Solutions (Context Assessment, Strategy Selection)

1. Review of literature for solutions others may have utilized for same or similar problems

2. Review of research in progress which may have bearing on problem

3. Review (as feasible) of various problem settings where "local solutions" may be evolving

D. Identification of General Program Approach (Strategy Selection)

II. General Curriculum Design (Program Planning/Design)

A. Development of Program Goals and Objectives

1. Statement of purpose (value loaded/non-measurable)

2. Statement of major goals (non-measurable)

3. Statement of major objectives (usually non-measurable)

B. Specify Program Parameters and Assumptions

1. Identify major constraints (geographical, conceptual, political, etc.)

2. Identify key assumptions (about learners, instructors, setting, etc.) which will be treated as "fact" for design purposes

C. Develop Learning Design

1. Identify major learning unit topics (purposes)

2. Develop learning unit objectives (state learner outcomes in measurable terms and specify rationale for unit)

3. Select the appropriate instructional strategy for each unit or sub-unit

4. Describe the teaching/learning activities for each unit or sub-unit

5. Describe the instructional materials needed for each unit or sub-unit

6. Finalize sequencing and time assignment for each unit or sub-unit

7. Prepare program summary

8. Prepare preliminary evaluation design: process and product objectives with timelines

III. Instructional Materials Development (Program Planning/Design; Implementation, Process, and Product Evaluations)

A. Non-Print (Tape/Slide-Film)

1. Establish evaluative criteria for acceptance

a. Scripts

b. Graphics

c. Photography

d. Audio

2. Scripts

a. Develop directions for script writer(s)

b. Scripts and storyboards developed

c. Approval and/or revisions

3. Graphics

a. Develop directions for artist(s)

b. Graphics developed

c. Approval and/or revisions

4. Photography (slides-film)

a. Develop directions for photographer(s)

b. Actors and settings arranged

c. Film/photos shot

d. Prints developed and edited

e. Approval and/or re-shooting, etc.

5. Audio

a. Develop directions for audio

b. Arrange for voices, studio, etc.

c. Tape

d. Approval and/or revisions

6. Synchronization

a. Integrate all components

b. Approval and/or revisions

B. Print

1. Individual and group materials

a. Develop instructions for writers

b. Prepare materials

c. Approve and/or revise (validate)

2. Participant's Manual (Workbook)

a. Selection of instructional materials to be included (or used as handouts)

b. Development of introduction, transitions, etc.

c. Compile, edit, and revise

d. Determine layout

e. Pilot test and revise

f. Determine graphics and cover design

g. Final approval

3. Instructor's Guide

a. Selection of instructional materials to be included

b. Write instructions to be used in presenting materials to participants

c. Write general instructions for instructor

d. Compile, edit, and revise

e. Determine layout

f. Pilot test and revise

g. Determine graphics and cover design

h. Final approval

IV. Evaluation Design (Planning; Implementation; Process and Products

A. Identify process for reviewing goals, objectives, and learner outcomes for measurability and for proper selection process.

B. Identify process of reviewing materials during development for content validity, reading level, and appropriateness to stated objectives.

C. Develop process and product objectives, to permit checking of ongoing and ultimate attainment of steps in program development.

D. Develop plan for pilot and field testing

1. Identification of pilot and field test population

2. Development of testing process

3. Development of evaluation instruments

4. Development of data processing and reporting plans

5. Specification of the review process for instruments (i.e., curricular validity)

V. Pilot and Field Testing (Implementation; Process and Product)

A. Arrange for pilot and field test participants

B. Arrange dates and locations

C. Prepare materials and equipment

D. Conduct field tests

E. Analyze data

F. Revise materials and program as appropriate

G. Finalize materials for production, packaging, and distribution

VI. Institutionalization

A. Set up policies and procedures for regularizing

B. Assign resources (trainers, equipment, space, etc.)

C. Assign participants

D. Schedule

In the foregoing, a systems approach to training program development was described. Webster defines "system" as: a set or arrangements of things so related or connected as to form a unity or organic whole; a regular orderly way of doing something; a set of facts, principles, rules, etc. classified or arranged in a regular, orderly form so as to show a logical plan linking the various

parts." A system can also be described as a set of interdependent elements and the relationships between those objects. Following this line of thought, an organizational or training system can be described as consisting of sets of people, objects such as materials and equipment, and activities that are related through interactions and dependencies.

A systems approach is, of course, not particularly new. But, for our way of thinking, the systems approach should include some additional considerations and finer specifications that promote the development of social action, innovative, and change efforts, such as training programs. These points pertain, in general, to the issues and procedures of evaluation.

1. Evaluation is implicit in the systems approach, but should be made explicit; evaluation facilitates decision-making.

2. The systems approach implies that the whole is greater than the sum of the parts (i.e., team or group approach is necessary rather than autonomous individual units).

3. On-going feedback is essential to the systems approach (i.e., process evaluation).

4. Performance objectives are a necessary part of a system to determine when goals have been met (i.e., process and product evaluation).

5. Program evaluation in the "EIS" Approach is mission (goal) accomplishment-oriented.

6. The "EIS" focuses upon group problem-solving procedures within the social-psychological context.

It is interesting that evaluation has not been specifically related to, and integrated into, the systems concept. In the foregoing description of a training or development project, evaluation was used throughout. In the following, the "Evaluation-In-Systems" Approach is described in regard to steps or levels of evaluation. Although the specific procedures may vary from project to project, all projects, (developmental applied research, social action) are encompassed in this Approach.

THE "EIS" APPROACH

The "Evaluation-In-Systems" Approach to the development of training programs is built upon the systems approach and recent developments in educational evaluation. This section describes the different levels or types of evaluation and the components in development of educational programs (e.g., training programs).

A. Evaluation Levels or Types.

For our purposes, we find it necessary to consider seven (7) levels or types of evaluation (or assessment). These are:

1. Needs Specification: surveying, identifying, and ordering the needs, or deficit conditions that "need correcting."

2. Context Assessment: identifying the key organizational and situational features of the "task setting" that influence planning, decision-making, and ultimate agency satisfaction with the product. The key elements in the system and relationships between them are delineated.

3. Strategy Selection: consideration is given to alternative programs and strategies that might overcome deficit condition or improve existing situation. Alternative approaches are considered and weighed relative to policies and philosophy of agency. This step leads to the identification of a "preferred solution strategy." Policy analysis is performed here.

4. Program Planning: the parameters impacting the type of program are identified; resources and operations required to bring about the "solution" are specified. At this level the evaluation design is prepared which will guide subsequent development and decision-making. All the elements and interrelationships comprising the system are taken into account in the preparation of the evaluation design.

5. Implementation Evaluation: as the program is launched, assessment is made to see how closely the actual situation and developments coincide with those planned. Unexpected problems are identified and corrective procedures taken. Evaluation procedures insure that a "coordinated and integrated arrangement of elements" (system) is in operation. In training programs, the actual materials development is commenced here. (After the training materials are developed, pilot and field testing efforts are commenced at this level.)

6. Process Evaluation: assessment of the program is performed on an ongoing and continuing basis to insure that the program is on-course. This is done by checking accomplishment of process or interim objectives.

7. Product Evaluation: Determination whether outcome or product objectives of the program have been accomplished along with suggestions for the next round of development.

Most evaluation writers have chosen to speak about four types or levels of evaluation: needs assessment, planning, process, and product. Typically, needs specification and context evaluation are considered to be the same thing or context analysis is omitted. Stragegy selection (policy analysis) and program planning are lumped together, as are planning and implementation.

We believe that speaking only to needs specification and not specifically to context analysis is terribly unfortunate because the social context is underemphasized. The failure to specify the social context is to overlook the fact that a social organization or system exists which must be modified to accept the new program. A new program, inevitably, introduces changes in the traditional roles and functioning of individuals and of the organization. Related to this consideration of the social context is the need to consider the existing organizational factors, policies, and procedures in program selection and planning. If the context analysis is properly performed, the important social-personal elements and interactions that affect program selection and planning, and later success will have been identified.

We believe, too, that strategy selection is different from program planning because strategy selection entails analysis of organizational policies and philosophies that determine the type of approach or strategy that will eventually be chosen. Once a

strategy for overcoming a deficit condition is selected, a program can be planned and designed.

Finally, the pooling of planning and implementation provides inadequate attention to the special needs of implementation. In the social action arena, the proverbial "slip between the cup and the lip" has become notorious. Implementation is THE key step in program development. The ultimate recipients and users of a new program are usually inadequately considered, with the result that costly educational programs are inadequately tried out and are stored in closets and warehouses.

Training programs are like field experiments. The evaluation design contains consideration of the critical parameters of the experiment. Implementation evaluation insures that the planned study is implemented, in that people are doing what they are supposed to be doing. Process evaluation insures that "experimental controls" are in effect and that extraneous and unexpected factors are identified and considered. Product evaluation compares the actual with planned accomplishments and investigates the influence of various factors on the outcome.

B. Components in Educational or Training Program Development.

We believe that six (6) components are typically involved in developing educational and/or training programs. These components are:

1. Instructional: Specification is made of what will be taught by somebody (e.g., trainer, counselor, teacher) to somebody

(e.g., trainee, client, pupil) in some manner. Determination is made here that existing materials or procedures are inadequate.

2. Materials Development: new or special materials must be developed from scratch, modified, or adapted for use in the educational/training program.

3. Staff Development: personnel to use the new materials and do the instruction must be trained; they must be led to accept a new role definition for themselves, to modify existing behaviors, and to identify with the program.

4. Community Involvement: the community and social context in which the program will be implemented must be involved in the planning stages and throughout the effort; they need to input into the program and receive feedback; community includes those who will be more or less directly involved by the effects of the program. COMMUNICATION is key here!

5. Space/Time Structures: special space and time considerations must be taken into account, such as in queuing situations (e.g., computer-assisted instructions, first-aid delivery systems, logistical supply problems).

6. Management Procedures: the procedures for coordinating and accomplishing all aspects of the program must be specified.

We see all education/training programs being comprised of, or involving, these six (6) components.

For each of the six components, an evaluation design is prepared in the program planning phase. The evaluation design consists of product (outcome) objectives and one or more process

objectives for each product objective. Process objectives are used to indicate the various steps that must be accomplished to reach a certain outcome or product objective. All objectives have measurable criteria to facilitate determination of accomplishment or non-accomplishment.

C. A Special Feature: Internal Evaluation and External Evaluation.

A special feature of the EIS Approach is the use of an external evaluator in concert with the internal evaluation function. The external evaluation role is somewhat similar to that of a fiscal auditor. The external evaluator provides special content and methodological expertise to abet that of the project staff. He provides an increment of quality assurance to the planning and evaluation effort by further insuring that important features of the project have been considered.

Also, the external evaluator provides a level of objectivity that is important not only to project personnel but subsequent efforts to disseminate the training program. In the case of the affirmative action program for the hiring of handicapped workers developed by ARBEC, Inc., the use of the external evaluator served to promote dissemination of the material in that some potential consumers were insistent upon meeting with and discussing program features with the external evaluator rather than the project staff.

A matrix showing the 42 cells of the EIS Approach is presented on the next page in Chart I with narrative to provide some integration.

Evaluation should be distinguished from assessment because

they are often used interchangeably and, sometimes, inappropriately. Strictly speaking, assessment as measuring is what the evaluator usually does. Evaluation refers to the act of judging. Evaluation is typically the domain and responsibility of the decision-maker. In the "EIS Approach," the evaluation function is shared by the contractor, the grantee, and the external evaluator. The evaluation design provides a mutually agreed upon set of objectives and criteria that facilitate group problem-solving and group decision-making.

Chart 1. Evaluation-In-Systems (EIS)
Approach to Training Program Development.

Evaluation Levels	Program Components					
	Instructional	Materials Development	Staff Development	Community Involvement	Space/Time Structures	Management
Needs Specification	(Identify deficit conditions; specify and prioritize "needs.")					
Context Assessment						
Strategy Selection (Policy Analysis)	(Specification of elements and interactions that make up the whole (system).)					
Program Planning/ Design	(Evaluation components.)	Design is prepared (finalized) at this level...all program				
Implementation Evaluation	(Assessment made when program is launched.)*					
Process Evaluation	(Assessment is made on ongoing basis; monitoring and feedback operations.)					
Product Evaluation	(Outcome assessment is made; determine whether products (outcome) objectives are met.)					

*In training programs, development work commences here. After program materials are completed, pilot testing, field testing, and ultimately institutionalization of the total program is commenced at this level.

1134

References:

1. Carleson, Robert D. and Lewis, James A. The Systems Analysis Workbook. Prentice Hall Inc. Inglewood Cliffs, N.J., 1973.

2. Doffert, Marion L. Systems and Models. The Generator (Social Context of Education, Division G) American Educational Research Association, Vol. 6, No. 2, 1976, pp. 15-22.

3. Dutton, Maurice G. (Editor). Employment of the Handicapped, Affirmative Action Training Program. ARBEC Inc. Austin, Tex., 1976. Vol. 5.

4. Liberty, Paul G. Texas Rehabilitation Commission Supervisors' Training Program, Evaluation Report. ARBEC Inc. Austin, Tex., 1974.

5. von Bertalanffy, Ludwig. General Systems Theory. New York: Braziller, 1968.

CHARACTERISTICS OF AN OFFICER PERFORMANCE EVALUATION MEASURE

by

Arthur C. F. Gilbert, Ph.D.
Frances C. Grafton

A Paper Prepared for Presentation at the
18th Annual Military Testing Association (MTA) Conference
October 18-22, 1976

Individual Training and Performance Research Laboratory
U.S. Army Research Institute for the Behavioral and Social Sciences
Arlington, Virginia 22209

1257

CHARACTERISTICS OF AN OFFICER PERFORMANCE EVALUATION MEASURE

Arthur C. F. Gilbert, Ph.D.
Frances C. Grafton

U.S. Army Research Institute for the Behavioral and Social Sciences[1]
Arlington, Virginia 22209

The leadership research program of the U. S. Army Research Institute for the Behavioral and Social Sciences (ARI) continues to focus on the dimensions of leadership and on the development of predictive indices of leadership. Of particular interest to the research described in this paper are the findings of Helme, Willemin, and Grafton (1971) who reported eight broad dimensions of leadership performance. These dimensions were defined on the basis of an analysis of officer behaviors in a realistic field experimental combat situation and consisted of the following factors: technical-managerial leadership, combat leadership, team leadership, command of men, mission persistence, executive direction, tactical staff skills, and technical staff skills.

As part of an ongoing research effort on the relationship between active duty performance and measures of officer performance and potential obtained during the Officer Basic Courses (OBC) it was felt essential to develop a criterion instrument that would reflect critical aspects of leadership as defined in the research of Helme, Willemin, and Grafton (1975) and from other available sources (Stogdill, 1974). This instrument the Performance Evaluation Form, was briefly described by Gilbert (1975) but will be described in greater detail in this paper.

---

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

The objectives of the research reported in this paper were (1) to evaluate the reliability of the Performance Evaluation Form and (2) to determine if the instrument could differentiate among the requirements of alternative officer assignments. In view of these objectives, it was hypothesized that differentiation among branches on the variables of the Performance Evaluation Form should be maximized if two branches that differ substantially in functional assignments were evaluated. Consequently, the Infantry Branch was selected as one that best represented the Combat Arms. The factor structure of ratings in this branch should emphasize the combat aspect of the variables in question. The Quartermaster Branch was selected as being a representative branch of the Services Support Branches. In this instance, the factor structure of variables should differ markedly in terms of ratings from those in the Infantry.

PROCEDURE

The Performance Evaluation Form. As described briefly elsewhere (Gilbert, 1975) the "Performance Evaluation Form" provides a measure of overall performance and nine ratings along dimensions that are considered as being crucial in the leadership domain. The ten scales of the Performance Evaluation Form are shown in Figure 1.

An assessment of the officer's overall performance is required in the first section of the evaluation instrument. A 7-step scale was adapted from Willemin (1965) (Figure 2); the high end "7" is described as "Outstanding" and the low end "1" as "Unsatisfactory".

Figure 1 shows the seven scales of the second section of the Performance Evaluation Form. For each scale the specific attribute that each scale purports to measure is given. Five of the seven scales correspond to the factors derived by Helme, Willemin, and Grafton (1971) while two correspond to the principal leadership dimensions derived from Ohio State University research (Stogdill, 1974). The rater was required to rank order these scales in terms of the officer's relative potential for future assignments and then he was required to rate the officer on the seven step scale as defined in Figure 2. In the third section of the Performance Evaluation Form the rater was required to rate the officer's potential performance in Combat Leadership and in Technical Managerial Leadership on the seven-step scale.

Data Collection

Officers in the 13 Career Branches and in the Chaplain Corps who attended the Officer Basic Courses (OBC) during the Fiscal year 1974 were used as subjects in the overall research program. Approximately 5,000 officers were involved in the overall research program. Four Performance Evaluation Forms were mailed for each officer in order to

Figure 1

PERFORMANCE EVALUATION FORM SCALES AND CORRESPONDING FACTORS

Performance Evaluation Form Scale	Factor
Part I	
Duty performance	
Part II	
Applying tactical knowledge and skills in support of combat operations	Tactical staff skills[1]
Understanding the mission and clearly defining personal roles of subordinates in its accomplishment	Team leadership[1]
Making decisions and initiating actions under pressure	Command of men[1]
Defining functional roles and duties in the process of developing subordinates to fill assignments for long-term unit effectiveness	Initiation of structure[2]
Planning and organizing manpower and materiel to meet situational requirements	Executive direction[1]
Motivating troops to accomplishing the mission by taking into consideration their well being morale	Consideration[2]
Applying knowledge of logistics and technical matters to solve support problems	Technical staff skills[1]
Part III	
Combat leadership	Combat leadership[1]
Technical-managerial leadership	Technical-managerial leadership[1]

[1] Helme, Willemin, and Grafton (1971)
[2] Stogdill (1974)

Figure 2

OFFICER PERFORMANCE SCALE[1]

Scale Value	Description
7.	OUTSTANDING: Far above the requirements of the situation, suggesting the highest kind of formal recognition through meritorious award, or decoration.
6.	SUPERIOR: Markedly above the requirements of the situation, suggesting formal recognition through a special (favorable) efficiency report, or letter of commendation.
5.	ABOVE AVERAGE: Somewhat above the requirements of the situation, suggesting informal recognition through specific favorable comment in his regular efficiency report, and through informal appreciation or commendation.
4.	AVERAGE: Fully up to the requirements of the situation, suggesting general appreciation (Perhaps mostly unexpressed).
3.	BELOW AVERAGE: Somewhat below the requirements of the situation, though suggesting only the mildest kind of corrective action through informal recommendation for improvement, or through change of duty assignment within the organization.
2.	MARGINAL: Markedly below the requirements of the situation, suggesting formal corrective action through a special (unfavorable) efficiency report, administrative admonition, letter or reprimand, summary court, or transfer out of the organization.
1.	UNSATISFACTORY: Far below the requirements of the situation, suggesting the most drastic kind of formal corrective action through reclassification, demotion, general court, or boarding out of the Army.

[1]Adapted from Willemin (1965)

obtain ratings from four different raters. The organization personnel officer was responsible for insuring that the Performance Evaluation Form was completed by each officer's immediate supervisor, a superior officer other than his immediate supervisor and two close associates. The three raters other than the immediate supervisor were designated by the personnel officer in consultation with the officer's immediate supervisor.

Four complete ratings were obtained on 2,886 officers. The distribution of these officers in the 13 Career Branches and in the Chaplain Corps is shown in Table 1.

Data Preparation

All records generated from the Performance Evaluation Forms were edited to insure that the rankings and ratings were compatible on the seven variables in the second section of the form. In other words, the rank of the officer's area of greatest potential should have been "1" and his weakest area "7". When this condition did not obtain the ratings or the rankings were corrected. For the research reported here the officers in the Infantry and in the Quartermaster branches were used as subjects. A total of 771 Infantry officers and 102 Quartermaster officers were involved in the analyses.

RESULTS AND DISCUSSION

The reliability estimates for the ten scales of the Performance Evaluation Form were computed by averaging the six possible correlations obtained among the four raters. The resulting average correlations were adjusted by use of the Spearman-Brown Prophecy Formula. The ten reliability estimates for the Performance Evaluation Form are shown in Table 2 for the Infantry Branch and for the Quartermaster Branch.

Examination of Table 2 indicates that for the Infantry Branch the reliability estimates ranged from .74 for the global rating of duty performance to .59 for the scale that describes the officer's potential ability to apply knowledge of logistics and technical matters to solve troop support problems. The reliability estimates for the Quartermaster Branch ranged from one of .70 for the scale that described global duty performance to .50 for both the scales that reflected the officer's potential to perform in the Combat Leadership domain and in the Technical-Managerial Leadership domain. Generally, the obtained reliabilities somewhat more in line with those reported by Willemin (1965) than those reported for the Ranger sample (Gilbert, 1975).

The obtained reliability estimates were evaluated to detect statistically significant differences in magnitude between the two branches.

Table 1

DISTRIBUTION OF OFFICERS BY BRANCH WITH
FOUR COMPLETE RATINGS

Branch	Number
Air Defense	22
Adjutant General	176
Armor	354
Chaplain	27
Engineer	205
Field Artillery	520
Finance	49
Infantry	771
Military Intelligence	34
Military Police	121
Ordnance	120
Quartermaster	102
Signal Corps	352
Transportation Corps	33
Total	2,886

Table 2

RELIABILITY ESTIMATES FOR THE TEN SCALES OF THE PERFORMANCE EVALUATION FORM FOR THE INFANTRY BRANCH AND FOR THE QUARTERMASTER BRANCH

Scale	Reliability Estimate Infantry N = 771	Quartermaster N = 102
Part I		
Duty performance	.74	.70
Part II		
Applying tactical knowledge and skills in support of combat operations	.67	.69
Understanding the mission and clearly defining personal roles of subordinates in its accomplishment	.63	.62
Making decision and initiating actions under pressure	.71	.63
Defining functional roles and duties in the process of developing subordinates to fill assignments for long-term unit effectiveness	.60	.64
Planning and organizing manpower and materiel to meet situational requirements	.63	.55
Motivating troops to accomplish the mission by taking into consideration their well being and morale	.65	.69
Applying knowledge of logistics and technical matters to solve troop support problems	.59	.56
Part III		
Combat Leadership	.71	.50
Technical-managerial leadership	.64	.50

Only one statistically significant difference appeared. The reliability estimate for the scale measuring potential to perform under combat leadership conditions was significantly higher for the Infantry Branch than it was for the Quartermaster Branch. This lower reliability estimate for the four raters in the Quartermaster Branch could very well be due to the fact that in that Branch there is not as much of an opportunity to assess this characteristic as there is in the Infantry Branch.

After the reliability estimates were obtained the data were analyzed to determine if the ratings of the variables reflecting potential performance in different areas could differentiate between the Infantry and Quartermaster branches and the extent to which the four ratings (i.e. immediate supervisor, other superior, and the two associates) might be different within the branches.

The first analysis involves computing the correlation among the nine variables across all the four raters for the Infantry Branch. This correlation matrix was then factor analyzed using a principal components solution using the highest absolute row value as the diagonal estimate. Four factors were extracted that had associated eigenvalues of at least equal to 1.00. These factors were then rotated using the varimax method. The loadings of the ratings on the nine scales of the Performance Evaluation Form on these four factors (raters) are shown in Table 3. At this point, it should be noted that throughout this paper, each of the nine variables in the second and third section of the Performance Evaluation Form will be referred to by the leadership behavior it is purported to measure (Figure 1).

As shown in Table 3 the loadings for the variables on each factor are high and uniform in magnitude on the four factors. These four factors then can best be described as each being representative of a particular type of rater but not yielding very much information about the way in which each rater individually rates individuals on the variables in question. Consequently, it was considered desirable to explore the factor structure of the ratings given by each type of rater to seek a solution to this problem. Since only rater factors emerged in the analysis of the Infantry data, it was felt that separate analyses for each rater in the Quartermaster Branch would also be appropriate.

A correlation matrix for each of the four raters in each branch was computed for the nine variables of the second and third section of the Performance Evaluation Form. Each of the eight matrices obtained was factor analyzed. A principal components solution was obtained using the highest row value as the diagonal estimate.

Table 3

LOADINGS OF RATINGS BY THE FOUR RATEES ON THE FOUR FACTORS IN THE OVERALL ANALYSIS OF INFANTRY RATINGS

Variable	IS*	OS*	A*	A*
1. Tactical Staff Skills	.77	.81	.78	.76
2. Team Leadership	.82	.85	.78	.81
3. Command of Men	.81	.83	.81	.81
4. Initiating Structure	.80	.80	.76	.78
5. Executive Direction	.82	.80	.78	.79
6. Consideration	.73	.76	.74	.74
7. Technical Staff Skills	.72	.73	.72	.75
8. Combat Leadership	.79	.82	.80	.77
9. Technical-Managerial Leadership	.76	.79	.79	.76

Note: IS = Immediate Supervisor
OS = Other Supervisor
A = Associate

In each of the eight solutions only one factor with an associated eigenvalue of 1.00 or greater emerged. The per cent of common variance associated with this first factor for each analysis is shown in Table 4. In Table 5, the unrotated factor loadings of the nine variables on this factor are shown for each analysis.

The loadings for each of the four raters on the nine variables for the Infantry Branch appear to be similar in magnitude. Across the four sets of loadings the higher loading appear for the ratings on the factors corresponding to command of men, team leadership, combat leadership and executive direction. More exact tests for differences among the factor loadings or among the raters for these loadings is risky because of the different magnitude of the eigenvalues obtained in each matrix and because of the interpendence among the loadings in a given matrix. Thus, any interpretation here has to be of an exploratory nature.

The loadings for each of the four raters on the nine variables for the Quartermaster Branch appear comparable. If one considers the four loadings for each of the variables, the three variables having the highest loadings are command of men, team leadership, and executive directions.

Comparison of the loadings of the variables on the first factor for the two branches did not indicate any marked trend. In other words, the loadings obtained from the four raters in the Infantry and in the Quartermaster did not appear to be different.

Generally, it appears that as far as ratings of performance potential on the variables are concerned, there is a certain uniformity within each branch as to how the ratee is perceived. Also, it appears that there is a certain uniformity across the two branches.

Lack of differentiation among the ratings on the variables for the two branches as hypothesized led to exploration of the differences that might exist for the rankings of these variables between the two branches. Consequently, the next series of analyses focused on the ranking of the officer's potential performance on the seven variables in the second section of the Performance Evaluation Form. These analyses paralled the analyses of the ratings except that, now, only seven variables were involved.

Again, correlation matrices for each of the four raters in each branch were computed for the seven variables in the second section of the Performance Evaluation Form. Each of these matrices were factor analyzed using a principal components solution using the highest absolute row value as the diagonal estimate.

Table 4

PERCENT OF VARIANCE CORRESPONDING TO THE FIRST FACTOR EXTRACTED IN THE FACTOR ANALYSES OF RATINGS

Analysis	Per Cent of Variance
Infantry:	
Immediate Superior	90.75
Other Superior Officer	91.71
Associate	91.17
Associate	90.22
Quartermaster:	
Immediate Superior	87.73
Other Superior Officer	87.55
Associate	85.39
Associate	86.24

Table 5

LOADINGS OF EACH VARIABLE ON THE FIRST UNROTATED FACTOR FOR EACH OF THE EIGHT ANALYSES OF RATINGS

Variable	Infantry				Quartermaster			
	IS*	OS*	A*	A*	IS*	OS*	A*	A*
1. Tactical Staff Skills	.82	.85	.81	.81	.68	.68.	.74	.77
2. Team Leadership	.87	.88	.84	.86	.81	.84	.84	.86
3. Command of Men	.86	.88	.86	.86	.83	.86	.83	.86
4. Initiating Structure	.83	.83	.82	.81	.81	.80	.83	.86
5. Executive Direction	.85	.84	.84	.83	.81	.82	.85	.80
6. Consideration	.79	.80	.80	.78	.68	.76	.75	.72
7. Technical Staff Skills	.74	.76	.78	.77	.66	.81	.75	.74
8. Combat Leadership	.84	.88	.86	.83	.78	.81	.81	.80
9. Technical-Managerial Leadership	.79	.83	.84	.79	.79	.89	.77	.74

*Note: IS = Immediate Supervisor
OS = Other Superior Officer
A = Associate

In each of these analyses, only one factor has an associated eigenvalue of 1.00 or better and in the analysis of the rankings of the superior officer in the Quartermaster Branch, the first factor had an associated eigenvalue of .9215. The per cent of common variance accounted for by the first factor in each analysis is shown in Table 6, while the unrotated factor loadings are shown in Table 7.

Examination of the loadings for each of the variables for the unrotated first factor solution for the Infantry Branch, again reveals a certain uniformity. In other words, within the Infantry Branch the four raters tended to perceive the ratee in similiar fashion. If the loadings of the variables are considered across the four raters a bipolar factor appears to emerge. One end of this factor is defined by command of men while the other end is defined by technical skills. Tactical skills and team leadership load substantially and positively on this factor when the four sets of loadings considered.

In the Quartermaster Branch, on the other hand, there seemed to be differences in the factor structure that emerged in each of the four analyses as revealed in Table 7. If the loadings across the four raters are considered as well as inspection of the individual loadings for each of the four raters, the first factor that emerges is defined by the variable corresponding to technical staff skills at one end and command of men at the other end. Executive direction loads positively on all four unrotated factors for this branch. Tactical skills loads negatively for three solutions.

In summary then, the factor analytic solution of the rankings appears to give a different factor structure for the Infantry Branch and for the Quartermaster Branch. Considerable uniformity appears to exist among raters in the Infantry Branch while this is not so within the Quartermaster Branch.

The results of this research indicate that there is not a different factor structure underlying the ratings in the Infantry Branch and in the Quartermaster Branch. This could very well be due to the high correlations among variables for a given rater. However, when the rankings of the variables for each rater are considered, there is evidence for differences in factor structure between each branch. The emergence of more clearly defined factors and differences between branches in the context of these analyses may be due in part to the fact that each attribute in the Performance Evaluation Form was described by a single statement.

Table 6

PER CENT OF VARIANCE CORRESPONDING TO THE FIRST FACTOR EXTRACTED IN THE EIGHT FACTOR ANALYSES OF RANKINGS

Analysis	Per Cent of Variance
Infantry:	
Immediate Superior	56.31
Other Superior Office	59.56
Associate	61.80
Associate	54.65
Quartermaster:	
Immediate Superior	47.32
Other Superior Officer	43.62
Associate	47.98
Associate	49.48

Table 7

LOADINGS OF EACH VARIABLE ON THE FIRST UNROTATED FACTOR FOR EACH OF THE EIGHT ANALYSES OF RANKINGS

Variable	Infantry				Quartermaster			
	IS*	OS*	A*	A*	IS*	OS*	A*	A*
1. Tactical Staff Skills	.46	.55	.58	.55	-.21	-.56	-.58	.25
2. Team Leadership	.30	.33	.35	.24	-.24	.01	.08	-.60
3. Command of Men	.59	.55	.58	.52	-.36	-.36	-.37	-.27
4. Initiating Structure	-.26	-.27	-.30	-.22	.28	.35	.02	-.34
5. Executive Direction	-.45	-.47	-.52	-.59	.53	.67	.59	-.17
6. Consideration	.08	.06	.04	.16	-.37	-.28	-.20	-.19
7. Technical Staff Skills	-.67	-.67	-.64	-.61	.44	.26	.50	.65

Note: IS = Immediate Supervisor
OS = Other Superior Officer
A = Associate

Factor loadings reflected for the two associate ratings in the Quartermaster Branch.

REFERENCES

Gilbert, A. C. F. Prediction of Army Officer Performance. Paper presented at the 17th Annual Conference of the Military Testing Association, Indianapolis, Indiana, September 15-19, 1975.

Helme, W. H., Willemin, L. P., & Grafton, F. C. Dimensions of leadership in a stimulated combat situation (Technical Research Report 1172). Arlington, Va.: U. S. Army Research Institute for the Behavioral and Social Sciences, July 1971. (NTIS No. AD-730 315)

Stogdill, R. M. Handbook of leadership; a survey of theory and research. New York: The Free Press, 1974.

Willemin, L. P. Criterion aspects of Army research on the prediction of officer performance (Research Study 65-6). Arlington, Va.: U. S. Army Research Institute for the Behavioral and Social Sciences, December 1965.

Performance Objectives: Problems in Definition Measurement, and Interpretation

Kenneth I. Epstein

Army Research Institute for the
Behavioral and Social Sciences

The first three steps in the Interservice Procedures for Instructional Systems Development (TRADOC Pamphlet 350-30, 1975; hereafter ISD) are I.1 Analyze Job, I.2 Select Tasks/Functions, and I.3 Construct Job Performance Measures. Part of the output of these first steps is a listing of all of the tasks selected for training along with a job performance measure (JPM) for each one. Each JPM includes the required test performance, conditions, cues, and standards, and for purposes of this paper will be considered equivalent to performance objectives. These JPM's become the fundamental basis for the development and control of training in the ISD process.

The role of objectives in ISD, and the major emphasis in most of the literature on objectives, is to describe as clearly and precisely as possible what students will be required to do after instruction. Thus, JPM's, which describe performance on the job, drive the development of the learning objectives used in training. In addition, objectives can form the basis for determining whether training is effective, whether individual students have successfully completed training, or, in some cases, whether training is necessary. The methodology suggested in ISD for developing JPM's implies that exhaustive task lists will be required along with detailed statements

The opinions expressed in this paper are those of the author and do not necessarily imply endorsement by the U. S. Army

of conditions and unambiguously defined standards for acceptable performance.

The purpose of this paper is to explore some of the problems and limitations in developing and using objectives beginning with the ISD methodology as a point of departure. Detailed listing of tasks and conditions will be discussed in terms of practicality and the trade off between precisely defining objectives and generalizing to a more comprehensive behavioral domain. Problems in defining standards for an objective and then translating objective standards into test performance criteria will be addressed. Finally, some of the problems in applying objectives as described in ISD to unit performance will be briefly discussed.

The first step in the ISD process involves a complete listing of all of the tasks thought to be required to perform successfully on the job for which training is being developed. From this list, those tasks which are critical for the job are separated from those which are redundant, extraneous, or "nice to know." The critical tasks are retained for training and JPM's are developed for each of them.

While this is certainly a desirable goal, problems can arise in trying to be extremely specific in defining tasks and in generating objectives for very large numbers of tasks. Examples of how extreme attempts to make objectives specific can be are found in Hively, Patterson, and Page (1968) and Millman (1974). For example, Hively, Patterson, and Page describe an objective for subtraction problems

that includes whether the problem should be written horizontally or vertically, the number and values of the digits included in the problem, and the placement of important steps in the subtraction process such as borrowing from the second or third place. In an address at the annual meeting of the American Educational Research Association, Popham (1976) described experiences with teachers trying to write objectives for their classrooms. He found that when teachers were required to develop exhaustive course objectives, they found it very difficult and tedious and generally produced poorly written objectives. On the other hand, when the emphasis was placed on a small set of "critical" objectives, the task was better received and acceptable objectives were produced. Popham, unfortunately, did not address the issue of determining which, of the set of all possible objectives, were the critical ones. The point is that it appears as if something less than a complete, highly specific listing is all that can be hoped for in developing sets of objectives.

Large numbers of tasks can also lead to testing problems. Often, resources are not sufficient for testing all tasks. In such cases, tasks must be sampled for testing and the performance on the sampled tasks generalized to the complete list. In order for such generalizations to be made, either assumptions or extensive empirical work is required to establish the interrelationships among the sampled and unsampled tasks.

Current research being conducted by the Army Research Institute (ARI), American Institutes for Research (AIR) and HumRRO has been

exploring the use of a quantitative approach to task selection. This approach uses cluster analysis and specially derived generalizability indices to generate qualification tests for tank weapons crew proficiency. The approach is detailed by Boycan and Rose (1976).

Each task statement includes some observable behavior. When the task is translated into an objective, this behavior is often referred to as the "action". ISD suggests that the action statement contain two parts; the first describes the type of learning that is to take place, the second describes how evidence that the learning occurred will be exhibited. The type of learning helps to determine what sorts of instructional design strategies are most likely to be effective. It also has implications for the type of evidence of learning that is desirable. For example, if the type of learning is a mental skill, something more than the rote repetition of the instructional material would be required. Otherwise acquisition of information, or the ability to state something, would be all that would be exhibited. Similarly, if the type of learning involved a physical skill, a verbal description of the skill in question would, at best, indirectly imply acquisition of the skill itself.

For any type of learning there may be more than one way to show that learning has occurred. For example, several learned facts could be recited orally or in writing. Acquisition of a concept (a mental skill) could be shown by choosing members of the concept class from a list, sorting a collection of cards into piles of members and nonmembers of the class, or providing an example of the concept. If

objectives are to be specific, then the particular form of evidence required must be included. Otherwise, equivalence of the possible forms of evidence must be, at least implicitly, assumed. For example, it is well known that the probability of guessing the correct answer to a multiple choice question is much greater than guessing the correct answer to the same question in constructed response format.

The necessity of a trade off between precision and generalizability is the issue here. Can performance on a multiple choice test be interpreted in the same way as performance on a constructed response test over the same material? Perhaps, but it is an empirical question and it may be answered differently for different types of learning and different content areas. Alternatively, is it possible to sacrifice precision and allow either type of test format by selecting test items randomly from a pool containing both? If so, how should the test results be interpreted? When practical considerations force a test to be inappropriate for the type of learning, the problem becomes worse. This is particularly a problem in tasks that require physical skills. Evidence for the ability to accomplish a primarily physical task, such as changing a tire, is often provided in the form of procedural lists or descriptions. Acceptance of such evidence requires rather strong assumptions about the relationship between the ability to describe a task and to do it. Whether such assumptions are valid or not is a largely unresolved question.

The conditions part of an objective "clearly states the limits and/or conditions of student performance. (ISD, Phase II, p. 13)." Conditions normally include the information provided to the student, the materials he is permitted to use, and the environment in which he will be required to perform the actions stated in the objective. For example, computational formulas may be provided, technical manuals may be available, calculators may be permitted, or other aids to accomplishing the required task may be permitted during testing. Environmental conditions could include constraints on time, weather, light or darkness, or the working environment in which the task is to be performed. Once again the problem of precision versus generalizability is present. If the requirement for precision is to be met, detailed specifications of the conditions are needed and inferences about performance must be restricted to performance only under the stated conditions. On the other hand, if the task is likely to be performed under a variety of conditions, and it is desirable to generalize to performance over all conditions, then some sort of sampling strategy, with its inherent lack of precision must be used. In some cases complete specification of conditions may be impossible. For example, a rifle marksmanship task may be performed on an outdoor range. To specify all of the possible conditions of visibility, temperature, wind direction and velocity, and other environmental conditions is an impossible task. Sometimes, however, an approximation is possible. For example, daylight vs night, clear weather or rain, or above and below fifty degrees. In any case, whenever performance under the particular conditions

which held during testing is accepted as the performance which would be exhibited under more general conditions, assumptions about the equivalence of conditions for task performance are, at least, implicitly made.

Cronbach, et. al. (1972) describe a methodology termed the theory of generalizability, for systematically studying the degree to which performance may be generalized within and across objectives and conditions. In order to apply the methodology, test scores are analyzed in a repeated measures analysis of variance with objectives and/or conditions determining the sources of variance. Thus, if objectives were homogeneous there would be relatively little variance due to objectives expected. Similar results would be expected for equivalent conditions. For the case where several objectives applied to very different skills or where varying conditions led to important differences in performance, these factors would be expected to account for a major portion of the variance observed. There is a need for generalizability studies to provide an empirical basis for determining the extent to which inferences about performance can be extended to a variety of objectives and conditions.

The standards part of an objective "is a clearly stated measurable standard of performance (ISD, Phase II, p. 14)." It is intended to provide a basis for determining whether or not an individual can perform the action described in the objective adequately. Considered as part of an objective, standards are usually idealizations. Normally they do not explicitly recognize psychometric problems that

might occur during testing. Nor are factors such as temporary manpower requirements which are external to the task but important for personnel decisions included in the standards part of most objectives. Standards will be considered first as idealizations. Secondly, standards will be considered as they are translated into actual test requirements.

Glaser and Klaus (1963) provide an excellent framework for conceptualizing standards. They suggest that, "In practice, proficiency standards can be established at any value between the point where the system will not perform at all and the point where any further contribution from the human component will not yield any increase in system performance (p. 424)." The Glaser and Klaus article represents one of the first applications of criterion-referenced measurement to performance evaluation. Notice that there is no mention of the performance of an individual relative to his peers. Rather, the standard is set relative to the actual requirements of the system in which an individual must operate. Clearly, some individuals may be able to perform at a higher level of proficiency than others, but this is irrelevant as long as system requirements are met.

The Glaser and Klaus framework is easiest to visualize in man-machine systems. For example, if the task involved vehicle maintenance, minimal standards could be determined in terms of the basic requirements for operation of the vehicle. If a trainee's eventual job will never call for extensive engine rebuilding, then the system

requirements imply some maximal performance standards below those that would be required for an engine overhaul. Similar limits on time can be imagined. For example, if four hours are available for vehicle maintenance, then four hours is the minimal standard of performance in terms of time. A two hour time requirement is not appropriate, and it would be wasteful to devote training time and money to try to achieve it. For other types of tasks which do not include man-machine interactions it is sometimes more difficult to specifiy standards. For example, the proportion of misspelled words that would constitute unacceptable performance in a task involving report writing is probably arbitrary. However, it should be possible to determine at what point poor spelling makes a report incomprehensible. While such standards may be, in some sense, arbitrary, it is important that they be specified and justified.

Assume, for purposes of argument, that standards stated in system requirement or other acceptable terms are available. The standards must be combined with the action and conditions portion of the objective to generate test items. Now assume further that the test developer is qualified and that the objective is well enough written that there is a good match between the objective and the test items. Depending on the task, the level of detail of the objective, and the type of learning involved, the test items and standards will be more or less explicit. For example, for procedural tasks which have objectives for each step in the procedure, the standard for each

objective may explicitly require perfect performance. That is, if each step in the procedure is not performed correctly, the system will not operate successfully. Other tasks may allow for more error in the human component. For example, since his work will be checked, a 5% error rate may be acceptable for a routine clerical task. For more academic subject matter, 10% or 20% error rates may still qualify as acceptable performance. In each of these cases the particular allowable errors may not be explicitly stated. Instead some general level of acceptable performance is specified. What do these standards imply for testing?

The type of learning has measurement implications. Gagne' and Briggs (1974) suggest some of these implications in their dicussions of appropriate learning strategies. Graham and Bergquist (1975) provide a detailed analysis of measurement implications of the mental skills and information types of learning. Briefly, two measurement models may be appropriate. The first considers the performance of a skill only as completely incorrect or completely correct. The second conceptualizes skill performance on a continuum from completely incorrect to completely correct. For example, Gagne' and Briggs define information as a collection of facts or an organized body of knowledge. At this level of detail it seems sensible to consider acquisition of information on a continuum without designating the specific facts that are known or unknown. Mental skills, such as applying mathematical rules, appear to be either acquired or not acquired in an all-or-none fashion. Thus, the "completely incorrect or correct" measurement model seems appropriate. If the "completely incorrect or correct" model is appropriate,

then the standards should theoretically require perfect performance. If the continuous model is appropriate, then the standards may be stated in terms of acceptable, but not necessarily perfect, performance.

Regardless of the idealized standards, factors such as imperfections in test items, unconrolled environmental conditions during testing, and a wide variety of psychological effects will lead to measurement error. Such error must be considered when operational standards are specified. For example, the subtraction skills necessary for keeping a running balance in a checkbook may have been well learned and normally exhibited as an accurate balance. However, even the best "subtractor" occasionally makes an error. Since subtraction is a mental skill which theoretically and empirically is acquired in an all-or-none fashion, such occasional errors must be considered errors of measurement. In the case of partial acquisition similar problems occur. For example, a task may require learning how to spell several hundred technical terms. The objective may require the ability to spell any 80% of those terms. If all the words could be tested, then it would probably be reasonable to require 80% accuracy and assume that any measurement error would be absorbed in the 20% allowable error rate. But, what if resources require that only a portion of the entire list be sampled for testing? How many words should be sampled, how should they be sampled, and how should standards be set? Current research at ARI is exploring these issues in the context of criterion-referenced measurement (CRM). A variety of CRM models have been reviewed and evaluated in terms of their

statistical and decision making characteristics, assumptions, and score interpretations (Macready, et.al, in press). Applications of CRM models in cases where item sampling is a problem are discussed by Steinheiser and Snyder (1976).

Since the results of testing will normally be used to make decisions, the impact of those decisions should be considered in setting standards. Most of the decisions relevant to this discussion of learning objectives reduce to dichotomous, pass-fail, go-nogo, acceptable-unacceptable judgments. For the case of dichotomous decisions two types of error can occur; capable persons can be judged unacceptable or incapable persons can be judged acceptable. These errors are called false negatives and false positives respectively. The costs of such errors will vary enormously depending on the particular task, the possibilities of remedial training, the demands for personnel in the field, and other factors which interact with the training system. Generally, if standards are high, the chances of committing false negative errors will be high relative to the chances of committing false positive errors. This is simply because high standards require superior performance for that performance to be acceptable. If standards are lowered, false negative errors will tend to decrease and false positive errors will tend to increase. Thus, if the cost of judging an incapable person acceptable is high, as would be the case in aircraft maintenance, one would like to keep the chances of committing a false positive error as low as possible and standards would be high. Conversely, if on-the-job training were

inexpensive relative to remedial school training, one might be willing to accept a relatively high false positive error rate and allow standards to fall.

Thus far the discussion has assumed that tasks could be defined, at least conceptually, in great enough detail to permit objectives to be written. Is this always the case? As more is learned about unit performance, particularly in a simulated combat environment, the answer to this question begins to look more and more negative. The reasons for this include the enormous variety of events that can occur in combat, the problems in describing how individuals interact to accomplish unit tasks, the overwhelmingly large number of conditions that impact on performance, the lack of readily quantifiable standards, and, perhaps most difficult, the ever changing dynamic nature of combat. The theory behind instructional objectives which was designed primarily with a school environment in mind does not seem to directly apply to unit performance in simulated combat. It provides a strong foundation, but new thinking and much new development is required. Johnson's (1976) paper describes a research effort at ARI, aimed at some new approaches to developing and using performance objectives for evaluation.

This paper has surfaced a multitude of questions without providing many answers. In fact, each question could probably lead to a paper. The purpose here was not to discredit or undermine instructional objectives. Quite the contrary, it is only the existence of

objectives that has allowed such questioning to occur. However, it is critical that the limitations of objectives be recognized and that research be continued to overcome those limitations.

References

Boycan, G. G. and Rose, A. M. Evaluation of crew gunnery performance Paper presented at the annual meeting of the Military Testing Association, Gulf Shores, AL. October, 1976.

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. The dependability of behavioral measurement: Theory of generalizability for scores and profiles. New York: John Wiley & Sons, Inc. 1972.

Gagné, R. M. and Briggs, L. J. Principles of instructional design. New York: Holt, Rinehart and Winston, Inc. 1974.

Glaser, R. and Klaus, D. J. Proficiency measurement: Assessing human performance in Gagné, R. M. (Ed.) Psychological principles in system development. New York: Holt, Rinehart and Winston, Inc. 1963.

Graham, D. L. and Bergquist, C. C. An examination of criterion-test characteristics in relation to assumptions about the nature of achievement variables. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C. April, 1975.

Hively, W., Patterson, H. L. and Page, S. H. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Johnson, E., III. Simulated combat as a test bed: The Army Training and Evaluation Program (ARTEP). Paper presented at the annual meeting of the Military Testing Association, Gulf Shores, AL. October, 1976.

Macready, G. B., Epstein, K. I., Steinheiser, F. H., Jr., and Mirabella, A. Methods and models for criterion-referenced testing. Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, in press.

Millman, J. Criterion-referenced measurement. In Popham, W. J. (Ed.) Evaluation in Education: Current applications. Berkeley, CA. McCutchan Publishing Corporation, 1974.

Popham, W. J. Expanding the technical base of criterion-referenced test development. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. April, 1976.

Steinheiser, F. H., Jr. and Snyder, C. W., Jr. Criterion-referenced test modeling for individual and weapon crew performance. Paper presented at the annual meeting of the Military Testing Association, Gulf Shores. AL. October, 1976.

U. S. Army Training and Doctrine Command. TRADOC Pamphlet 350-30, Interservice procedures for instructional systems development. Fort Monroe, VA., 1 August, 1975.

SCORE QUALITY ISSUES RELATED TO INDIVIDUAL AND WEAPON CREW CRITERION-REFERENCED PERFORMANCE TESTS

Frederick Steinheiser, Jr.
Army Research Institute for the Behavioral and Social Sciences, Arlington, Virginia

and

C. Wesley Snyder, Jr.
American Institutes for Research
Washington, D.C.

Presented at the
Military Testing Association Conference
October 1976

1287

Presented at the Military Testing Association Conference, October 1976.*

Score Quality Issues Related to Individual and Weapon Crew Criterion-Referenced Performance Tests

Frederick Steinheiser, Jr.
Army Research Institute for the
Behavioral and Social Sciences
Arlington, Virginia

and

C. Wesley Snyder, Jr.
American Institutes for Research
Washington, D.C.

In criterion-referenced personnel testing, an examinee is advanced if he is able to respond correctly to at least some specified percentage of the items or trials comprising the test domain. This minimum passing percentage, or "criterion level," reflects the degree of mastery deemed sufficient to pass a specific instructional or training objective. The actual percentage of items that he would correctly respond to in the population of items may be termed his "true level of functioning" (or "true state of combat readiness"). In practice, the advancement-retention decision must be made from a fairly small sample of items. As a result, errors in the classification of examinees are unavoidable. The task of the examiner is to minimize these errors as much as possible through good test design.

Due to a variety of practical constraints, such as time, manpower, and cost of testing, test length is an important consideration. A reasonable goal for the decisionmaker is to be able to classify examinees into groups of masters and nonmasters (or combat ready vs. not combat ready) as accurately as possible, without requiring excessive numbers of test items or trials. This problem is addressed by the first two models presented in this paper: (i) the binomial model, which provides the probability of an

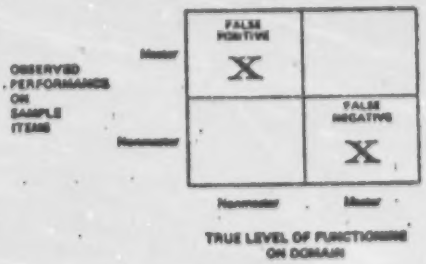
*The views expressed in this paper are those of the authors and do not necessarily imply endorsement by the U.S. Army.

examinee obtaining a test score given his assumed true level of functioning; and (ii) the Bayesian model, which provides a probabilistic estimate of an examinee's true level of functioning given his test score.

Another practical constraint in personnel evaluation concerns the heterogeneity of items on different tests, and the heterogeneity of various examinee populations. Especially in military testing, it may not be possible to give each examinee group the same set of items or trials. The third section of this paper presents a model which allows examinees of differing skill levels to be scaled on a common skill metric, even when they have not all taken the same set of test items.

Binomial Model

A marksman or tank crew unit is assumed to have a true level of functioning (i.e., when tested on the entire domain, the examinee would "pass" a certain proportion of items). Based on a hypothetical standard involving some proportion of items passed, we can split the examinees into two groups--masters and nonmasters. When actually tested on a subset of the domain (e.g., 18 items) with a specified mastery level (e.g., 83 percent), we can identify masters (e.g., those who pass 15 out of 18 items) and nonmasters. The situation is diagrammed below.



In each of the four boxes an inference can be drawn about the true level of functioning given performance on a sample of items or about performance, given true skill. The two boxes marked with an "X" indicate errors on the part of the examiner. First, he might pass examinees whose true level of functioning is below the specified mastery level (called a false positive error). Alternatively, he might fail examinees whose true level of functioning is at or above the specified mastery level (called a false negative error). The probability of these errors occurring is a function of the number of items included in the test. The binomial model serves as an analytic aid to the consideration of the test length issue in terms of

probable misclassification, provided the decisionmaker is willing to specify true skill levels for the intended examinee population.

In order to calculate the expected error in decisionmaking, we need to specify the lowest proficiency level required for an examinee to be considered a master (called minimal proficiency level) and the highest proficiency level that an examinee could attain and still be considered a nonmaster (called maximum nonproficient level). These levels would intuitively be at the same point except that the test is not completely accurate. Given these parameters as set by the decisionmaker based on his knowledge of the situation, the probability of false positive and false negative errors for minimal masters and maximal nonmasters can be calculated for any given test length and cutting score.

For example, assume that a true proficiency (which cannot be directly observed) of 90 percent is set as the minimal proficiency level, and that a true proficiency of 70 percent is set as the maximum nonproficient (not ready for combat) level. The region between these two cutoff scores is an area of uncertainty. Perhaps if an examinee's score fell in this region, he would receive a brief course for refresher training, whereas if his score were 70 percent, he would receive extensive retraining.

Continuing with this example, we also need to specify values for classification errors. Suppose that the examiner is unwilling to accept more than 25 percent of the examinees whose true proficiency state is 70 percent; and he wants to reject no more than 25 percent of those whose true proficiency is at 90 percent. Thus, the probabilities for false positive and false negative misclassifications have been set before any data have been collected, at .25 each. Given the values of lower bound for true proficiency (90 percent), upper bound for true non-proficiency (70 percent), and tolerable false positive and false negative error rates (.25 each), it is then possible to determine the minimum number of trials that should be given and how many trials should be correct. Table 1 shows, for example, that for an observed score of 80 percent, at least nine trials must be given and at least eight "hits" must be made in order to evaluate a crew as combat ready, subject to the above constraints. (The computational details may be found in an Army Research Institute paper by Macready, Epstein, Steinheiser, and Mirabella, 1976.)

Table 1. Probability of misclassification* as a function of the length of Table VIII.

Required Mastery Level	Table VIII Length	Passing Score	True Level of Tank Crew Functioning				
			50%	60%	70%	80%	90%
	36	29	--	1	11	43	2
80%	18	15	--	3	16	50	10
	9	8	2	7	20	56	23
	36	26	1	9	53	9	--
70%	18	13	5	21	47	13	1
	9	7	9	23	54	26	5
	36	22	12	48	9	--	--
60%	18	11	24	44	14	2	--
	9	6	25	52	27	9	1

* Entries to the left of dotted lines are probabilities (percent) of false positives; those to the right of dotted lines are false negatives.

(Adapted from Millman, 1973 and Novick and Lewis, 1974)

Another way of using the binomial model is to consider error rates for different cutoff points and lengths of an existing test. The present Army Table VIII for Tank Gunnery Qualification consists of 18 items. Inspection of Table 1 indicates that for an 80 percent cutoff level, this test length decreases the misclassifications for the 70 percent and 90 percent level to .16 and .10, respectively. Doubling the test to 36 items further reduces the misclassification to .11 and .02, respectively. The decisionmaker can assess the trade-offs of test-length and misclassifications in terms of the practical constraints of his test situation in order to arrive at an optimal test design.

It is obvious that in any evaluation program, some people who are qualified will be classified as not (yet) qualified, and vice versa. There is no such thing as giving an examinee population a set of 10, 100, or even 1,000 trials and hoping to be able to exactly separate the "sheep from the goats." But this binomial model quantifies the worst level of misclassifications that might be expected for any observed proficiency level, when test length and cutoff scores have been specified by the examiner.

Bayesian Model

Whereas the binomial model provides us with information about potentially observable behavior based on the decisionmaker's best guesses about the true skill levels in the intended examinee population, the Bayesian model improves the classification of observed performance. Turning it to the advantage of the test developer, the Bayesian model can be used to evaluate the usefulness of what is already known about the examinee population to improve the situation outlined by the binomial model.

Consider a testing program which classifies examinees into three combat readiness (proficiency) states, quantifiable as percentages: 90 percent indicates combat ready, 80 percent indicates a need for brief refresher training, and 70 percent indicates a need for extensive retraining. Suppose that the probability of getting a tank crew with a 90 percent readiness state is thought to be .60. This means that on the basis of historical information, 60 percent of all tested examinees has been found to have a readiness state of 90 percent. Also suppose that the probability of observing an 80 percent proficient crew is 30 percent, and the probability

of observing a 70 percent proficient crew is 10 percent. (The examiner's prior information about the proficiency distribution of the examinee population is represented by .6, .3, and .1.)

If we observe that an examinee scores 13 hits out of 18 trials, what is the probability that this examinee is combat ready, or needs brief refresher training, or needs extensive retraining? Note that on the basis of the observed data, 13/18 = 72 percent, which would seem to indicate a high probability for need of extensive retraining.

The test data can be combined with the prior information in the following way. The sampling distribution of the number of hits in 18 trials, for any proficiency, may conveniently be assumed to be a binomial distribution. It is then possible to go to a table of binomial probabilities to find the likelihood of getting 13 hits in 18 attempts with a 90 percent chance of success; 13 out of 18 with an 80 percent chance of success; and 13 out of 18 with a 70 percent chance for success. The values for these likelihoods are shown in column 3 of Table 2.

Next, we multiply the prior probability (column 2) times the likelihood value for each row. These values are shown in column 4. The sum of these values in column 4 equals .0785. The column 5 (posterior) probabilities are obtained by dividing each column 4 row entry by the column 4 total. The entries in column 5 represent the probability that an examinee is at a certain (hypothetical) state of readiness, having observed his test score and taking the prior expectations into account.

Thus, there is about a 17 percent chance that this crew is combat ready; about a 58 percent chance that they may need some short-term refresher training; and a bit more than a 25 percent chance that they may need extensive retraining. Note that the raw data of 13 hits out of 18 attempts yield a proficiency state of 72 percent. However, when the prior information that only 10 percent of all examinees can be expected to need extensive retraining is used, it becomes more likely that only short-term retraining is needed. Perhaps this is an "almost combat ready" crew that just had a "bad day," rather than being truly non-proficient. Of course, it is ultimately up to the decisionmaker or examiner to interpret these data.

Table 2. Bayesian computation of posterior probabilities for hypothetical states of readiness when test length, test score, and prior probabilities are specified.

1 Readiness State, R_x	2 Prior Probability of Observing R_x	3 Likelihood* (h= no. hits, n = no. trials)	4 Prior Times Likelihood	5 Posterior Probability, or $p(R_x/\text{test score})$
$R_1 = .90$	.60	$p(h=13/n=18, R=.90) = .0218$	.0131	.167
$R_2 = .80$	.30	$p(h=13/n=18, R=.80) = .1507$	.0452	.576
$R_3 = .70$	.10	$p(h=13/n=18, R=.70) = .2017$	.0202	.257
Total	1.00		.0785	1.00

* These values can be found in Hays, W.L. and Winkler, R.L., Statistics (Volume 1), New York: Holt, Rinehart, and Winston, Inc., 1970, Table V, p. 612.

The central characteristic of this model is its ability to use prior information. The more accurate the prior information, the greater the classification accuracy. Further, fewer test trials need be given in order for the examiner to reach a given level of confidence in making classifications. This latter point is very important for weapons crew evaluation, where each live-fire round may cost in the neighborhood of $100. The interaction of test length and quality of prior information is discussed more fully in Novick and Lewis (1974) and Epstein and Steinheiser (1975).

Comparison of Binomial and Bayesian Models

These models are similar in that they take the number of test trials into consideration. Consider a baseball analogy: a pinch hitter may bat only 10 times during a season, yet get three hits. In contrast, some players are able to end the season with an average of about .300, based upon hundreds of times at bat. Clearly, an index of "percent success" must be sensitive to the number of trials!

The binomial model is comparable to classical hypothesis testing procedures and is able to provide estimates for classification errors when the proficiencies, test length, and cutoff scores are specified. Or, it can prescribe a test length to meet the requirements of pre-set acceptable levels of both kinds of classification errors. For example, we could build utilities into the formulation that reflect, among other things, the type and magnitude of classification error for which we are willing to settle. False positive crews may be more acceptable than false negative crews at Fort Hood, but the opposite might be the case in West Germany or Korea, where we might want false positives to be near zero.

The Bayesian model, in contrast, cannot give misclassification probabilities. But it does give the probability of each examinee being at a certain proficiency level. Perhaps more importantly, it is able to use long-term, historical, "prior" information, and can thus exploit the data base which has evolved over time (years, perhaps) about the score distribution of previously tested examinees. To the extent that this prior information is accurate and stable, then classification accuracy will be enhanced, and the number of test trials to reach a specified level of accuracy will be reduced. However, inaccurate prior information can also

potentially degrade classification accuracy and may require a larger number of trials to reach a specified level of classification accuracy.

Considering the usefulness of each approach, an appealing alternative would be to use both for test development. Since performance tests are frequently expensive to trial run, particularly for large sample sizes, the use of analytic procedures which have minimal or no data requirements is very desirable. To proceed, the models can be sequenced in the order of their data requirements. Starting with the classical binomial model, the test developer can estimate the required dimensions of the test within the practical constraints of the situation. If good prior information exists, then the Bayesian model can provide more precise information (Novick & Lewis, 1974) and consequent savings in test time and equipment costs.

The application of these models requires attention to the following considerations (Novick & Lewis, 1974). First, the mastery level must be set, either by doctrine or by expert concensus, so that it reflects realistic expectations of performance after training. As the level approaches complete proficiency, the test length must be increased to provide accurate information for classification. Second, the tolerable levels of misclassification must be determined. Of course, the less error tolerance, the more resources necessary for the measurement process and the longer the master test must be. Third, the Bayesian model requires good prior information which can be translated into probability terms. Given stable conditions and quality historical data, the test can be shortened without loss in accuracy. Also, it is more realistic to think of proficiency not as a set of point estimates, but as regions along a continuum. Thus, we may want to say that all examinees who were 90 percent or more proficient should be classified as combat ready; those with proficiencies less than 90 percent but greater than 70 percent should be tested again; and those with proficiencies less than 70 percent should be retrained. These developments are discussed in a paper by Epstein & Steinheiser (1976).

Rasch Model

The model which we shall now consider does not concern itself with minimal numbers of items to make evaluative decisions. Rather, the Rasch

model calibrates items so that we can make statements concerning the relative skills of a set of examinees. This can be accomplished by giving one tank crew a test composed of trials a, b, c, d, e, and another a test of trials a, c, e, f, g, where each letter indicates a different distance or target type. All that is required is that the tests given to examinees have some items in common. The purpose of this section is to show how this logistic model can be used to equate tests (align the metrics of tests with items in common), and how such use may increase the utility of the the measurements.

The latent trait model developed by Rasch (1960, 1961, 1966) claims to yield person-free test calibrations and item-free person measurements. That is, the resulting scale is independent of the sample used to develop it, and the person "skill-score" is independent of the particular subset of items used in the test. The significance of the Rasch model may be appreciated by comparing it with "classical" models of test development (Whitely & Dawis, 1974, 163-164):

> A psychological test having these general characteristics would become directly analogous to a yardstick that measures the length of objects. That is, the intervals on the yardstick are independent of the length of the objects and the length of individual objects is interpretable without respect to which particular yardstick is used. In contrast, tests developed according to the classical model have neither characteristic. The score obtained by a person is not interpretable without referring to both some norm group and the particular test form used. . . .

The advantage of using the Rasch model in test development is that:

> . . . .No longer would equivalent forms need to be carefully developed, since measurement is instrument independent and any two subsets of the calibrated item pool could be used as alternative instruments. Similarly, independence of measurement from a particular population distribution implies that tests can be used for persons dissimilar from the standardization population without the necessity of collecting new norms.

These are very important and practical objectives to achieve, because in many military testing situations it is often difficult, if not impossible, to test all groups of examinees with the same test items. A specific example involves the evaluation of tank gunnery crews on both U.S. and European ranges with different Table VIII's.

Aligning the metrics of tests which have items in common can be easily accommodated by the model (K fer, 1976). Figure 1 depicts three Table VIII's given to three different tank units.

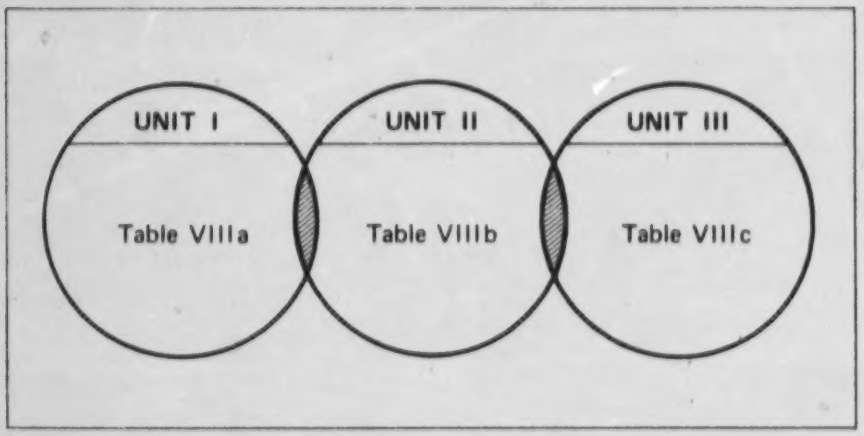


Figure 1. Three Table VIII tests (a, b, and c) with overlapping items given to three tank units (I, II, and III).

As depicted, units I and III are given some items that are in common to Unit II. What we want to be able to do is to compare all three units on the same metric, even though Unit I and Unit III have no items in common.

The items of the Table VIII's would represent some unidimensional aspect of tank gunnery. As indicated in the Boycan and Rose paper (1976), we can conceptualize the tank gunnery domain as eight task families, of which four are main gun tasks. Suppose tank units I, II, and III are tested on proficiency of the tank gunner to fire the main gun in the precision mode (Family V). Using the Rasch model, we would calibrate the items of each of the tests separately. This can be accomplished using a computer program (Mead & Wright, 1976), which is on file at the Army Research Institute computer center. The program produces maximum likelihood estimates of item difficulty and statistics for testing the fit between the Rasch model and the item data.

Item calibration helps us identify problem items. Typically, we examine the reliability and validity indices of trial items in order to determine the final set for measurement. The Rasch model is a form of

item analysis, but the emphasis is on the compatibility of the items with the model. Items fail to fit the Rasch scale if the model proves to be too simple to adequately characterize the performance data on the items or if the items measure skills other than the intended skill (Wright & Panchapakesan, 1969). In the case of our tank gunnery example, once we eliminate the unsatisfactory items and recalibrate on the final set, we would have three separate, revised tests, each with items which fit the Rasch model.

Suppose that Unit I had 18 items in their test, with 6 of these overlapping with items taken by Unit II; Unit II had 12 items, with 3 of these overlapping with items taken by Unit III; and Unit III had taken a total of 6 items. The next step in the alignment of the metrics is to estimate an "equating constant" in terms of the Rasch item difficulty indices. Since all three tests are calibrated using the Rasch model, the resulting item difficulty estimates should differ only by a constant. As Kifer (1976) points out, this constant is the difference between the average difficulties for the overlapping items for each group. After identifying the overlapping items that fit the model, we compute the mean difficulties for each of the scales. We then select the metric in which we will express the scores; that is, we could use Unit I as our anchor group for our new scale, or Unit II, or Unit III, or we could define a new metric. Once selected, we calculate the differences among the average difficulties and add the constant to the item difficulties for the groups that require adjustment.

In the tank crew application, we can express the scores on the Unit II metric. If the overlapping six items had the following Rasch item difficulty distributions for units I and II

Item	Unit I	Unit II
a	.53	.43
c	.29	.19
f	.46	.36
h	.41	.31
i	.60	.50
k	.37	.27

then we can easily see that the equating constant in terms of Unit II is -.10 (Avg. Difficulty for Unit I = .44 and the Avg. Difficulty for Unit

II = .34) and that the relationship between the two distributions is perfect (correlation of 1.00). More realistic data are given in Figure 2, where we plot the Rasch item difficulty estimates for Unit I versus Unit II.

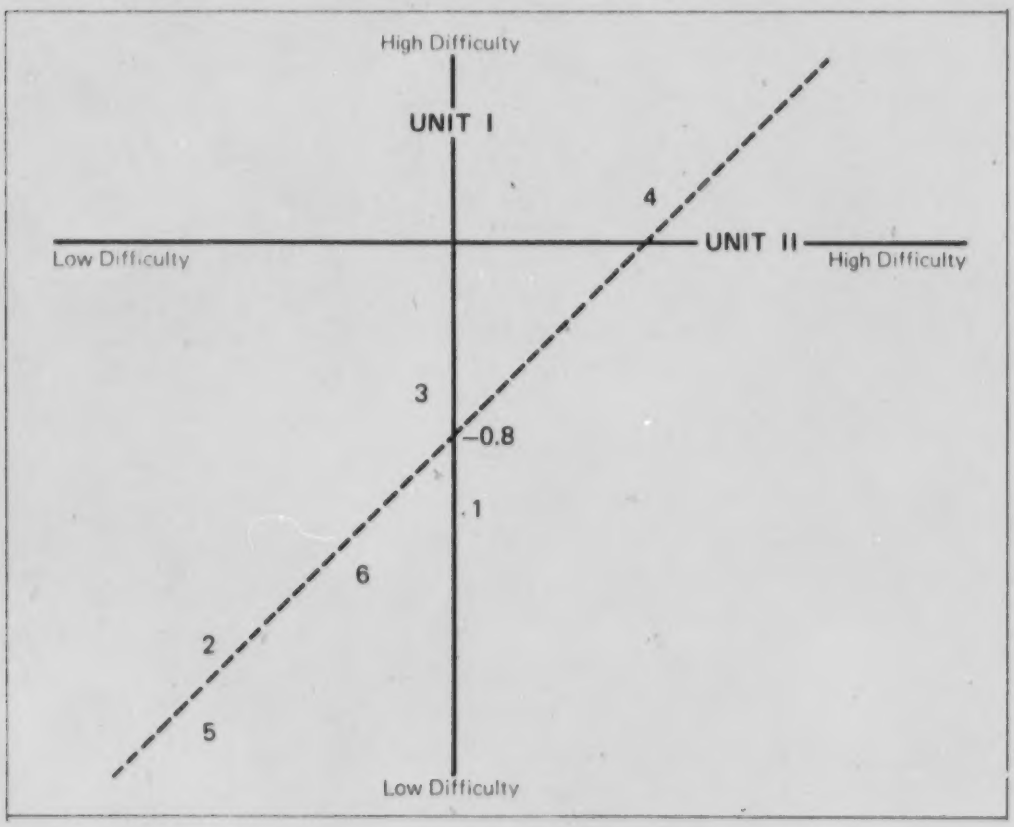


Figure 2. Plot of the Rasch Item Difficulties for Unit I versus Unit II.

If the item selection procedure has worked well, then the relationship will resemble that of the previous example. Although this fit is not so exact as the first, we have depicted a reasonable dispersion. This calls attention to an important consideration in the use of the Rasch model. Although it does not yield estimates of optimal test length, the precision of the model to estimate proficiency depends primarily upon the number of items used in the analysis. As indicated by Wright & Panchapakesan (1969, p. 26), "it is possible to reach any desired level of precision by varying the number of items used in the measurement, just providing that the range of item easiness is reasonably appropriate to the abilities being measured."

The equating constant for Unit I, calculated from the data in Figure 2, is .80. A similar analysis for Unit III might yield a constant of -1.20.* To complete the alignment we add .80 to the Unit I difficulties and -1.20 to the Unit III difficulties. We then re-estimate the skill levels and generate a new scoring table (computer scoring routine is available from Kifer). The result of the alignment of metrics is to increase the skill scale of Unit I and to lower the scale of Unit III. Proficiency scores for units I, II, and III are now on a common scale, and differences can be compared directly.

The virtues of the Rasch model can be summarized as: (1) Once a test has been calibrated on any group of examinees, it can be given again to a different group, without the need to create "parallel forms." For example, a test could be given to tank crews at Fort Hood for whom it had been developed originally, and later given at other ranges in the United States and Europe. (2) All proficiency scores will be on the same scale, regardless of the subset of items from which those skill levels were estimated. Thus, crew (examinee) A could be given a "hard" test and crew (examinee) B an "easy" test, and their scores would still be directly comparable. (3) Based upon the results of item calibration, one could take any subset of the items and express proficiency estimates on the common metric.

Some drawbacks are: (1) The original pool of items may have 20 percent items which do not fit the model and, therefore, several dozen items and several hundred examinees are suggested for the initial calibration of items. (2) Quantitative estimates of misclassification probabilities are not given as output; optimal cutoff scores for a given number of items are not derived.

Conclusions and Applications to Current Testing Programs.

The three approaches outlined above provide different kinds of information relating to the development of a test and use different amounts of information to arrive at their solutions. All relate to the quality of the score that will result. The binomial model yields the probability that

*See Kifer (1976) for an example of the technique with three science achievement tests on three student age groups in Sweden.

logistic model discussed in the last section of this paper will also make a contribution to the evaluative process, but in a different manner than the first two models. This model would allow scores to be placed on the same "gunnery" metric, even though the tests were conducted using different ranges.

References

Boycan, G. & Rose, A. An analytic approach to estimating the generalizability of crew performance objectives. Paper presented at the annual meeting of the Military Testing Association, Gulf Shores, Alabama, Oct. 1976.

Epstein, K.I., & Steinheiser, F.H. Comparison of error rates and misclassification probabilities using binomial and Bayesian models for personnel classification. Paper presented at the 22nd Conference on the Design of Experiments in the Army Research, Development and Testing, held at Harry Diamond Laboratories, 20-22 Oct. 1976.

_____. A Bayesian method for evaluating trainee proficiency. In 8th Naval Training Equipment Center/Industry Conference Proceedings. NTEC: Orlando, Florida, 19-20 Nov. 1975.

Hays, W.L. & Winkler, R.L. Statistics (Volume 1). New York: Holt, Rinehart, & Winston, 1970.

Kifer, E. Estimating scores on a common metric using the Rasch model: An I.E.A. application. Paper presented at the annual A.E.R.A. meeting, San Francisco, April 1976.

Macready, G.B., Epstein, K.I., Steinheiser, F.H., & Mirabella, A. Methods and models for criterion-referenced testing. Army Research Institute, Technical Paper, in press, 1976.

Mead, R., & Wright, B. Using the Rasch model for analyzing behavioral data. Interim Report to the Army Research Institute from the University of Chicago, Nov. 1976.

Millman, J. Passing scores and test lengths for domain-referenced tests. Review of Educational Research, 1973, 43, 205-215.

Novick, M., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Center for the Study of Evaluation Monograph Series in Evaluation, III: Problems in criterion-referenced measurement. Los Angeles, Calif: Center for the Study of Evaluation at U.C.L.A., 1974.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche, 1960 (English translation).

_____. On general laws and the meaning of measurement in psychology. In Proceedings of the fourth Berkeley Symposium on mathematical statistics. Berkeley: University of California Press, 1961, IV, 321-334.

_____. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

examinees will obtain a certain score given a hypothesized "true" level of performance. It provides an initial approximation of test length and cutoff scores without test data. The Bayesian model yields the probability that a particular examinee is a member of a certain proficiency group given a specific score. In this case, prior information on typical performance is combined with the binomial model information to relate the observed score to a true performance level. We therefore require information about the examinee population before the scores are observed. The model can improve the classification accuracy of the test and help us set cutoff points and fix test length at a more efficient level. The Rasch model yields the probability that an examinee in a particular skill group will answer a particular item correctly, given the easiness of the item. We need a good deal of information to accomplish the item calibration and person measurement associated with the Rasch model. Here the emphasis is on the particular items that we will rely on for the measurement; the final set is comprised of items which fit the model.

The Army Research Institute is using these models to analyze test data from two sources; individual pistol marksmanship and tank crew gunnery skills. In the first case, Military Police cadets will fire a .45 caliber pistol at stationary targets, at varying distances and positions. In this rather large experiment, A.R.I. obtained the cooperation of the Military Police School at Fort McClellan, Alabama. Three hundred cadets will fire a total of 240 rounds; this will assure a rich data source for input to the models. In return, A.R.I. will be able to provide specific information to the Military Police School about optimal cutting scores for specific length tests, anticipated misclassification rates, and interpretation of scores once they have been put on a common marksmanship metric. In the second case, the hit-miss scores of 653 tank crews firing main gun rounds through CONUS and USAREUR according to Table VIII are being analyzed. The binomial and Bayesian models are appropriate to the tank gunnery data, because each trial (firing a main gun round) is quite costly. The task of the decisionmaker is to make evaluative (i.e., "pass-fail," or "combat ready" vs. "not combat ready") judgments on the basis of test score data, using as few trials as possible. Subject to these constraints, the binomial and Bayesian models will tell the decisionmaker what his chances are of making correct and incorrect classifications. The

Whitely, S.E., & Dawis, R.V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-168.

Wright, B.D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

EVALUATION OF CREW GUNNERY PERFORMANCE

by

G. Gary Boycan
Army Research Institute
and
Andrew M. Rose
American Institutes for Research

INTRODUCTION

The Army program for gunnery training and evaluation of tank crews is defined by gunnery tables, which consist of several training or testing exercises where crews engage targets under various conditions. Crews progress through the exercises in Tables I through VII on their way to qualification: successful completion of the gunnery exercises in Table VIII.

The gunnery tables are developed by experienced Armor personnel who distill the essence of tank gunnery into a manageable set of training and testing exercises. Gunnery tables are developed for each Armor weapon system that has significant differences in firing capability and all tables are continuously revised to reflect changes in equipment and doctrine. Additional impetus for revising the gunnery tables is provided by the Army's commitment to increased combat readiness and increased efficiency in training and evaluation.

Efforts to improve the efficiency of tank crew training and evaluation are extensive. In addition to ongoing TRADOC and FORSCOM studies, a long range program of research has been initiated by the U.S. Army Research Institute. Numerous issues are being investigated including methods of defining and measuring the performance domain of tank gunnery, the use of simulation devices as a replacement or complement to live firing, and the type of practice schedule required to develop and maintain required levels of proficiency. As part of this programmatic effort, a project is underway to develop a more efficient Table VIII test for M60A1AOS crews. The first of a series of studies toward this goal described performance objectives in the gunnery domain in terms of their potential generalizability: the extent to which performance on one item predicts performance on others.

Generalizability is an important consideration in tank crew training and evaluation. The gunnery performance domain which tank crews must master is considerable. A recent study reported by Kraemer, Boldovici and Boycan (1976) resulted in a list of 225 performance objectives, which defined all possible ways that targets could be

The opinions expressed in this paper are those of the authors and do not necessarily imply endorsement by the U. S. Army

neutralized with M60A1AOS weapons. Each objective represents a unique set of procedures to be applied by the tank crew, the environmental and tactical conditions under which such procedures were appropriate and the standards of speed and accuracy required. However, given the practical constraints of time, range facilities and costs associated with live ammunition, crews cannot be provided the opportunity to practice and subsequently demonstrate proficiency on all such job objectives. Therefore, an optimal subset must be identified. In the case of evaluation this subset should consist of those objectives which minimize the risk of misclassification - that is, either passing crews who are really not qualified, or failing crews who are qualified.

Traditional empirical item-analysis methods were not feasible for this project because of the resources required to obtain repeated measures on all tasks in the gunnery domain. In lieu of establishing item generalizability empirically with livefire data, an alternative analytic approach was used.

The validity of the approach rests on the assumption that the more task elements or behavioral steps that any performance objective has in common with other objectives, the greater the communality among those objectives. Furthermore, the greater the communality, the greater the probability that performance on the one objective is predictive of performance on the others. For example, suppose that it could be demonstrated that firing the main gun using the precision method has more task elements in common with firing the main gun battlesighted than with firing the coaxial machine gun. The assumption is that because of this greater communality, the precision main gun task is more predictive of the battlesighted main gun task than of coaxial machine gun performance.

TECHNICAL APPROACH

The basic approach used to establish communality among performance objectives in the tank gunnery domain is shown in Figure 1. Objectives were listed down the left column of a large matrix. Task elements were then listed across the top. Examples of task elements are, "Loader announces 'up'," "Gunner indexes HEP" and "Gunner levels bubble." A '1' or '0' was then entered under each task-element in the row for each objective, indicating that performance of the objective did or did not require performance of each of the task elements. The matrix was completed in this manner for 240 objectives and a total of 113 behavioral elements.

Certain "system state" assumptions had to be made early in the analysis in order to achieve consistency in specifying which task elements were and were not included in each objective. If the weapons were assumed to be loaded, for example, then the task elements involved in a target engagement would be different from those that would be involved if the weapons were assumed to be unloaded.

Behavioral Element / Object Number	1	2	3	4	5	6	7	8	9	10	11	12	. . .	n
1	1	1	1	1	1	0	0	0	0	0	0	0		
2	1	1	1	1	1	1	1	1	0	0	0	0		
3	1	1	1	1	0	0	0	0	1	1	1	1		
.														
.														
.														
240														

Figure 1. Concept of the data matrix.

Cluster Analysis of the Objectives

The first question in analyzing the 240 performance objective by 113 task-element matrix was whether subsets or "families" of objectives could be identified that had many task elements in common. If relatively homogeneous families of objectives could be identified, then the problem of assessing generalizability of each objective in the entire domain could be simplified. Instead of estimating generalizability of each objective to 240 objectives, generalizability could be determined for a much smaller set of objectives. Representation of the entire domain of gunnery performance objectives could then be insured by including the most generalizable objective(s) from each family.

Two cluster analyses were performed in an attempt to identify families of objectives. Both analyses began by calculating the "behavioral distance" between every pair of objectives. Many distance measures have been reported in the literature, but for the one-zero data in the objective-element matrix most of these measures are equivalent. The simple matching coefficient (SMC) was used to measure behavioral distance in the present analyses. The SMC measures distance by the proportion of elements which are identical between each pair of objectives. Thus, for two objectives which have exactly the same values on 20 out of the 113 elements, the inter-objective distance is 20/113 or .177.

Use of the SMC produces a matrix that shows the behavioral distance between every pair of objectives. Objectives that are "close together" in behavioral distance form the clusters of objectives. The process is amalgamative, in that the two closest objectives form the seed for the

first cluster. Nearby objectives are incorporated into this cluster until an objective is found which is too far away; this objective then forms the seed of a new cluster.

Two clustering algorithms which employ the SMC were considered. One of these, the Average Distance amalgamation algorithm, requires an assumption that the 113 behavioral dimensions are independent. Since this assumption seemed questionable, both this algorithm and another, the Direct Clustering algorithm (Dixon, 1975; Hartigan, 1972) which does not require the independence assumption, were used. The Direct Clustering algorithm was used first, to test the independence assumption. The more typical Average Distance algorithm was then used as a check.

Results of the Cluster Analysis

The Direct solution indicated dependencies among some elements but not to such a degree as to rule out the Average Distance algorithm. Both solutions were therefore used as a pair of converging operations. The Direct cluster analysis yielded 26 distinct clusters while the Average Distance solution recast the 240 objectives into 34 separate clusters. To examine the similarity or degree of overlap between the two solutions, Venn diagrams were drawn which indicated how each job objective had been categorized by the alternative solutions. An example of the Venn diagram is presented in Figure 2. Each circle represents a set of objectives clustered by one or the other analytic procedure. Clusters defined by the Direct solution are shown by solid lines. Clusters emerging from the Average Distance approach are shown by dotted lines. The numbers within clusters represent specific performance objectives.

Inspecting the Venn diagrams permitted identifying those clusters produced by the two solutions which overlapped or were interconnected. Eight cases were observed in which higher-order clusters existed, containing a mix of clusters from each solution, and having no overlap with other higher-order clusters. These larger sets were referred to as families of performance objectives.

Table 1 summarizes the eight families which were isolated in terms of their component clusters and objectives. The most compelling feature of Table 1 is the descriptive information provided for each family. These descriptions were obtained by examining each objective within a family to establish what it had in common with the other members. The consistency of the objectives within a family is all the more striking when one recalls that: 1) description was attempted only after the families had been identified; and 2) the descriptors which were used had not been included in the cluster analyses.

As a further check on the integrity of the families, each was examined in terms of its behavioral elements. A typical outcome is shown in Table 2 for Family V - Gunner, Main Gun, Precision, HEP/BEE. Table 2

Table 1. SUMMARY OF FAMILY STRUCTURE

Family	Number of Clusters	Number of Objectives	DESCRIPTION			
			Crew Member	Weapon	Firing Mode	Ammunition
I	2	13	Gunner	Coax	Card	762
II	5	20	Tank Commander	.50 Caliber	Nonprecision	50
III	6	22	Tank Commander	Coax	Nonprecision/Precision	762
IV	6	24	Tank Commander	Main Gun	Battlesight/Precision	All Main Gun
V	6	16	Gunner	Main Gun	Precision	HEP/BEE
VI	15	78	Gunner	Coax	Nonprecision/Precision	762
VII	10	31	Tank Commander Tank Commander Gunner	Coax Main Gun Main Gun	Range Card Range Card Range Card	762 All Main Gun All Main Gun
VIII	10	36	Gunner	Main Gun	Battlesight/Precision	SB/HT

1188

1309

1310

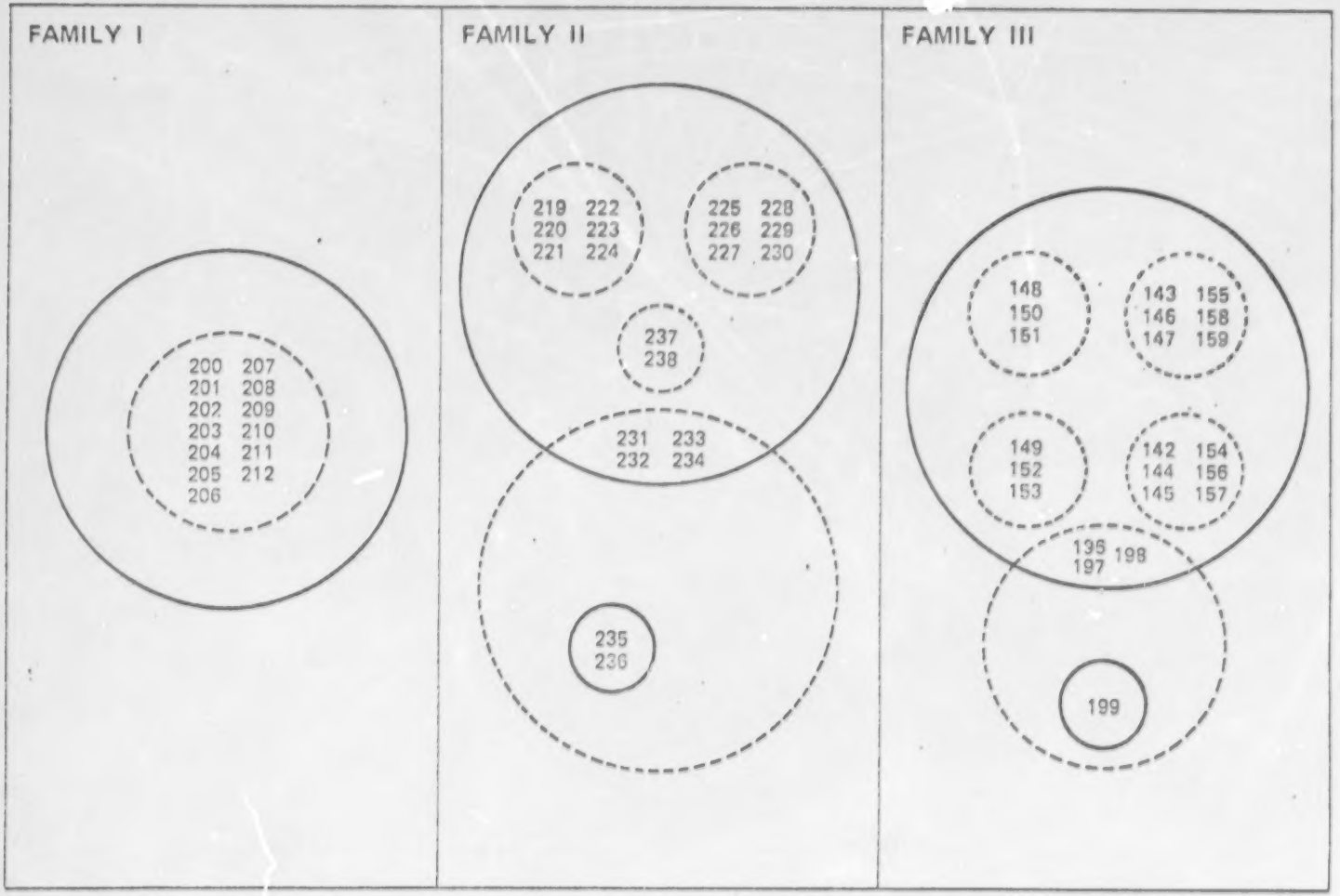
has two particularly salient features. First, only 39 of the total set of 113 elements are involved in Family V. Second, of these 39, 17 are possessed by every objective in the family. This patterning is typical of that found in all eight families: only selected sets of elements apply to each family, and many of these elements are possessed by every objective in the family. The latter elements primarily, but not completely, define a given family and distinguish it from others.

Based upon these results the strategy for selecting objectives for inclusion in a tank gunnery test would be to sample from the homogeneous objectives comprising each family. Since homogeneity is a highly relative concept, however, it can be argued that even greater homogeneity could be secured by subdividing families further on purely rational grounds. In Table 2 for example, one might further distinguish among the Family V objectives in terms of those involving HEP and those involving BEEHIVE ammunition. When this is done, two subfamilies can be identified. The first, containing objectives 49-56, contains 34 elements, 20 of which are common to all eight HEP objectives. The second or BEEHIVE subfamily, consisting of objectives 57-64, contains 35 elements, 22 of which are common to all 8 objectives. Given the outcome of the cluster analyses, the next step was to determine item generalizability.

Generalizability Analysis

The approach to estimating generalizability was based on the assumption that the more elements an objective had in common with other objectives, the greater would be the probability that performance on the one objective would be predictive of performance on the others. Two analogous indexes were developed to describe the generalizability of a given objective. One was used to express the generalizability of objectives within a given subfamily. The other index served a similar function at the family level.

The rationale for developing the generalizability index stemmed from the assumption that an objective consisting of elements occurring infrequently in the family or subfamily would not be as good a predictor of performance in that family as would another objective containing elements shared by many of the other objectives in the family. Since each objective has many elements the contribution of each element to generalizability was weighted by its frequency of family occurrence (F) or frequency of subfamily occurrence (f). One can, however, look at these same elements in larger perspective. That is, the elements also distribute themselves across the domain, some cutting across many different families, while others are concentrated almost entirely within a given family. These latter elements serve to define the family, and a parameter which takes this family-domain tradeoff into account was developed: F/D or f/D, where D is the frequency with which a given behavioral element occurred in the 240 objectives comprising the gunnery



1190

1313

1313

Figure 2. Venn diagrams defining families I, II, and III.

Table 2

Family V Objective-Element Matrix

1611 Description		17	18	24	25	26	27	28	29	30	34	35	42	44	45	46	47	48	51	52	53	56	57	60	61	64	66	67	68	69	70	71	14	15	16	35	37	38	43	45
		TC Announces Gnr	TC Announces Gnr, Whitelight	Driver Maintains Steady Speed	Driver Maneuvers Tank for Firing	Driver Announces Adverse Terrain	Driver Moves to Hull Down	Driver Brings Tank to Halt	Driver Locks Brakes	Loader Unlocks Ammo Ready Rack	TC Announces HEP	TC Announces BEEHIVE Time	TC Lays Gun for Direction	Loader Selects HEP	Loader Selects BEEHIVE	Loader Unloads Main Gun	Loader Loads Main Gun	Loader Places Main Gun Safety in Fire	Loader Announces HEP, Up	Loader Announces BEEHIVE, Up	Gunner Turns on Main Gun Switch	Gunner Indexes HEP or Coax	Gunner Indexes BEEHIVE	Gunner Selects HEP Reticle	TC Announces Target Description	Gunner Announces Identified	TC Announces Driver, Stop	TC Turns Off Computer	TC Ranges	TC Turns on Computer	TC Announces Range	Loader Sets BEEHIVE Fuze	Gnr Lays Crosshair at Center of Target	Gnr Lays Rangeline at Center of Target	Gnr Applies Aim-Off	TC Announces Fire	Gunner Relays on Target	Gunner Makes Final Precise Lay	Gunner Announces On the Way	Gunner Fires Main Gun
MG PRE S S B/C VI 5-32 D GPD HEP	49	X							X	X	X		X	X		X	X	X	X		X	X		X			X	X	X	X		X		X	X	X	X	X	X
MG PRE S S B/C VI 5-32 D TEL HEP	50	X							X	X	X		X	X		X	X	X	X		X	X		X			X	X	X	X			X		X	X	X	X	X
MG PRE M S B/C VI 5-32 D GPD HEP	51	X		X	X	X	X	X	X	X	X		X	X		X	X	X	X		X	X		X		X	X	X	X	X		X		X	X	X	X	X	X
MG PRE M S B/C VI 5-32 D TEL HEP	52	X		X	X	X	X	X	X	X	X		X	X		X	X	X	X		X	X		X		X	X	X	X	X			X		X	X	X	X	X
MG PRE S S B/C VA 5-32 N GPD HEP	53		X						X	X	X		X	X		X	X	X	X		X	X		X			X	X	X	X		X		X	X	X	X	X	X
MG PRE S S B/C VA 5-32 N TEL HEP	54		X						X	X	X		X	X		X	X	X	X		X	X		X			X	X	X	X			X		X	X	X	X	X
MG PRE M S B/C VA 5-32 N GPD HEP	55		X	X	X	X	X	X	X	X	X		X	X		X	X	X	X		X	X		X		X	X	X	X	X		X		X	X	X	X	X	X
MG PRE M S B/C VA 5-32 N TEL HEP	56		X	X	X	X	X	X	X	X	X		X	X		X	X	X	X		X	X		X			X	X	X	X			X		X	X	X	X	X
MG PRE S S TRP VI 5-32 D GPD BEE	57	X							X	X		X	X		X	X	X	X		X	X		X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
MG PRE S S TRP VI 5-32 D TEL BEE	58	X							X	X		X	X		X	X	X	X		X	X		X	X	X	X	X	X	X	X	X		X		X	X	X	X	X
MG PRE M S TRP VI 5-32 D GPD BEE	59	X		X	X	X	X	X	X	X		X	X		X	X	X	X		X	X		X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
MG PRE M S TRP VI 5-32 D TEL BEE	60	X		X	X	X	X	X	X	X		X	X		X	X	X	X		X	X		X	X	X	X	X	X	X	X	X		X		X	X	X	X	X
MG PRE S S TRP VA 5-32 N GPD BEE	61		X						X	X		X	X		X	X	X	X		X	X		X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
MG PRE S S TRP VA 5-32 N TEL BEE	62		X						X	X		X	X		X	X	X	X		X	X		X	X	X	X	X	X	X	X	X		X		X	X	X	X	X
S TRP VA 5-32 N GPD BEE	63		X	X	X	X	X	X	X	X		X	X		X	X	X	X		X	X		X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
S TRP VA 5-32 N TEL BEE	64		X	X	X	X	X	X	X	X		X	X		X	X	X	X		X	X		X	X	X	X	X	X	X	X	X		X		X	X	X	X	X

domain. Each element is weighted multiplicatively for family frequency and inversely for its frequency in the domain relative to the family (e.g., F . F/D). More simply stated, elements contribute to generalizability within a family to the extent that they are relatively common in the family and relatively unique in the domain.

Index values were computed for each objective in a given family as follows. First, for each behavioral element comprising the family the F^2/D or f^2/D value was computed. For a given objective, these values were summed across the elements defining the objective.

Results of the Generalizability Analysis

Generalizability index values represented estimates of the extent to which an objective permits generalizations about the family or subfamily of objectives from which it is drawn. A sample of the results is presented in Tables 3-5. Each Table represents one of the eight families and, when appropriate, constituent subfamilies. In each case, the code number and description of a performance objective in terms of 11 system variables is presented along with computed indexes of generalizability for the family and subfamily, if appropriate. Generalizability indexes are expressed in terms of z score equivalents and have been used to order objectives from most to least generalizable.

SUMMARY AND CONCLUSIONS

Generalizability is an important consideration in item selection, if test performance is expected to be predictive of performance on a larger domain. In this study, analytic means were used to estimate the generalizability of performance objectives which were candidates for inclusion in a tank gunnery test. Cluster analyses were conducted to identify subsets or families of objectives in the domain which were homogeneous in terms of behavioral elements. Then indexes of generalizability were computed to describe how generalizable a given objective was to others in its family. These analyses provided a useful framework for applying the criterion of generalizability during item selection.

The data base suggests including at least one objective from each of the subdomains that were identified. Given relatively limited resources, for example, the most generalizable objective from each of the eight major families would be included in the gunnery test. Given fewer constraints, the most generalizable objective from each of the 17 family or constituent subfamilies would be selected. While multiples of eight or 17 objectives represent the ideal, a partial sampling across families or subfamilies is also possible, and may even become necessary when other factors such as criticality, existing range facilities and ammunition costs are considered.

TABLE 3

Family 1 Objectives Ordered by Generalizability

ID NR	GUNNERY OBJECTIVE											
	CRW MEM	WPN	FIR MOD	VEH MOT	TGT MOT	TGT TYP	TGT VIS	TGT RNG	DAY/ NGT	F/C INS	AMMO	$\Sigma F^2/D$**
207	GNR	CX	RCL	STA	MOV	LAV	VAL	5-9	NGT	TEL	762	.67
205	GNR	CX	RCL	STA	MOV	LAV	VAL	5-9	NGT	GPD	762	.60
208	GNR	CX	RCL	STA	MOV	LAV	VAL	5-9	NGT	INF	762	.51
201	GNR	CX	RCL	STA	STA	L/C	VAL	5-9	NGT	GPD	762	.48
203	GNR	CX	RCL	STA	STA	L/C	VAL	5-9	NGT	TEL	762	.48
204	GNR	CX	RCL	STA	STA	L/C	VAL	5-9	NGT	INF	762	.45
209	GNR	CX	RCL	STA	STA	TRP	VAL	5-9	NGT	GPD	762	.28
211	GNR	CX	RCL	STA	STA	TRP	VAL	5-9	NGT	TEL	762	.28
212	GNR	CX	RCL	STA	STA	TRP	VAL	5-9	NGT	INF	762	.25
206	GNR	CX	RCL	STA	MOV	LAV	VAL	5-9	NGT	GPI	762	-.15
202	GNR	CX	RCL	STA	STA	L/C	VAL	5-9	NGT	GPI	762	-.27
210	GNR	CX	RCL	STA	STA	TRP	VAL	5-9	NGT	GPI	762	-.47
200	GNR	CX	RC	STA	STA	LCT	NVS	<900	D/N	AUX	762	-3.12

**Generalizability index values expressed as z-score equivalents.

TABLE 4

Family II Objectives Ordered by Generalizability

ID NR	GUNNERY OBJECTIVE											
	CRW MEM	WPN	FIR MOD	VEH MOT	TGT MOT	TGT TYP	TGT VIS	TGT RNG	DAY/ NGT	F/C INS	AMMO	$\Sigma F^2/D$
230	TC	50	NP	MOV	MOV	LAV	VAL	<1100	NGT	TPI	50	1.23
229	TC	50	NP	MOV	MOV	LAV	VAL	<1100	NGT	TPD	50	1.16
226	TC	50	NP	MOV	STA	LAV	VIS	<2300	D/N	TPD	50	1.04
227	TC	50	NP	STA	MOV	LAV	VAL	<1100	NGT	TPD	50	.96
228	TC	50	NP	STA	MOV	LAV	VAL	<1100	NGT	TPI	50	.78
225	TC	50	NP	STA	MOV	LAV	VIS	<2300	D/N	TPD	50	.59
224	TC	50	NP	MOV	STA	L/C	VAL	<1100	NGT	TPI	50	.58
223	TC	50	NP	MOV	STA	L/C	VAL	<1100	NGT	TPD	50	.52
222	TC	50	NP	STA	STA	L/C	VAL	<1100	NGT	TPI	50	.14
238	TC	50	NP	MOV	STA	AIR	VIS	<2300	DAY	TPD	50	.11
221	TC	50	NP	STA	STA	L/C	VAL	<1100	NGT	TPD	50	.07
236	TC	50	NP	MOV	STA	TRP	VAL	<1100	NGT	TPI	50	-.10
235	TC	50	NP	MOV	STA	TRP	VAL	<1100	NGT	TPD	50	-.17
232	TC	50	NP	MOV	STA	TRP	VAL	<1100	NGT	TPD	50	-.29
237	TC	50	NP	STA	MOV	AIR	VIS	<2300	D/N	TPD	50	-.33
234	TC	50	NP	STA	STA	TRP	VAL	<1100	NGT	TPI	50	-.42
233	TC	50	NP	STA	STA	TRP	VAL	<1100	NGT	TPD	50	-.49
231	TC	50	NP	STA	STA	TRP	VIS	<2300	D/N	TPD	50	-.61
220	TC	50	NP	MOV	STA	L/C	VIS	<2300	D/N	TPD	50	-2.17
219	TC	50	NP	STA	STA	L/C	VIS	<2300	D/N	TPD	50	-2.61

TABLE 5

Family III Objectives Ordered by Generalizability within Subfamilies

Subfamily III a

ID NR	CRW MEM	WPN	FIR MOD	VEH MOT	TGT MOT	TGT TYP	TGT VIS	TGT RNG	DAY/ NGT	F/C INS	AMMO	$\Sigma f^2/D$	$\Sigma F^2/D$
						GUNNERY OBJECTIVE							
153	TC	CX	NP	MOV	MOV	LAV	VAL	<500	NGT	RFI	762	1.58	1.63
152	TC	CX	NP	MOV	MOV	LAV	VAL	<500	NGT	RFD	762	1.26	1.41
149	TC	CX	NP	MOV	MOV	LAV	VIS	<500	D/N	RFD	762	1.26	1.40
147	TC	CX	NP	MOV	STA	L/C	VAL	<500	NGT	RFI	762	.98	1.05
146	TC	CX	NP	MOV	STA	L/C	VAL	<500	NGT	RFD	762	.66	.82
143	TC	CX	NP	MOV	STA	L/C	VIS	<500	D/N	RFD	762	.66	.82
151	TC	CX	NP	STA	MOV	LAV	VAL	<500	NGT	RFI	762	.53	.62
150	TC	CX	NP	STA	MOV	LAV	VAL	<500	NGT	RFD	762	.21	.39
148	TC	CX	NP	STA	MOV	LAV	VIS	<500	D/N	RFD	762	.21	.38
159	TC	CX	NP	MOV	STA	TRP	VAL	<500	NGT	RFI	762	-.08	.01
145	TC	CX	NP	STA	STA	L/C	VAL	<500	NGT	RFI	762	-.34	.21
158	TC	CX	NP	MOV	STA	TRP	VAL	<500	NGT	RFD	762	-.40	-.22
155	TC	CX	NP	MOV	STA	TRP	VIS	<500	D/N	RFD	762	-.40	-.22
144	TC	CX	NP	STA	STA	L/C	VAL	<500	NGT	RFD	762	-.66	-.01
142	TC	CX	NP	STA	STA	L/C	VIS	<500	D/N	RFD	762	-.66	-.02
157	TC	CX	NP	STA	STA	TRP	VAL	<500	NGT	RFI	762	-1.39	-.83
156	TC	CX	NP	STA	STA	TRP	VAL	<500	NGT	RFD	762	-1.71	-1.05
154	TC	CX	NP	STA	STA	TRP	VIS	<500	D/N	RFD	762	-1.72	-1.06
						Subfamiliy III b							
199	TC	CX	PRE	STA	STA	TRP	VAL	5-9	NGT	RFD	762	.87	-1.85
198	TC	CX	PRE	STA	STA	TRP	VIS	5-9	D/N	RFD	762	.86	-1.86
197	TC	CX	PRE	STA	STA	L/C	VAL	5-9	NGT	RFD	762	-.86	-.81
196	TC	CX	PRE	STA	STA	L/C	VIS	5-9	D/N	RFD	762	-.87	-.82

Use of the methodological approach just described need not be limited to tank crew performance objectives. Test designers working with any well defined performance domain may also find the approach to be a useful and convenient way of treating the issue of generalizability. In the near future, the U. S. Army Research Institute will publish a comprehensive report on its efforts to develop a more efficient test of tank gunnery. In addition to this study a number of related efforts will also be described. These include studies to establish the criticality of crew performance objectives, an assessment of the appropriateness of different test scoring procedures and a discussion of the tradeoffs in generalizability, criticality and score quality which are involved in the design of an efficient test of tank crew weapons proficiency.

REFERENCE

Kraemer, R. E., Boldovici, J. A. and Boycan, G. G. Job Objectives for M60A1AOS Tank Gunnery. Research Memorandum 76-9. U. S. Army Institute for the Behavioral and Social Sciences. April, 1976.

EVALUATION OF JUNIOR LEADERSHIP ASSESSMENT PROCEDURES

by

Captain Earle H. Bain

and

Captain John B. Joaquin

The views and opinions expressed are those of the authors and not necessarily those of the Department of National Defence

Presented to the 18th Annual Conference of the Military Testing Association at Gulf State Park Resort, Gulf Shores, Alabama, 22 October 1976.

CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
1107 Avenue Road
Toronto, Ontario, Canada
M5N 2E4

ABSTRACT

The Canadian Forces Personnel Applied Research Unit was requested by Training Systems Headquarters to evaluate current junior leadership assessment procedures prior to this training becoming a prerequisite for promotion.

The study investigated four specific questions:

(1) Are the 11 Critical Requirements (CRs) of leadership the instructors assess course participants upon, indeed the best/most important ones?

(2) Are peer ratings a valid form of leadership assessment in the junior leader school context?

(3) Are the Performance Objectives being assessed in a standard manner across the Junior Leader Schools?

(4) How is previous on job performance related to training results?

Analyses of two Study Questionnaires completed by 79 leadership school staff and 947 students found two critical requirements of leadership effectiveness (Assuming Responsibility and Command and Self-Assertion) were not being assessed!

Pearson's correlations of PEER rating leadership assessments and those of the instructors across seven different schools ranged from .38 to .60.

Comparing the mean total Critical Requirement scores and other course outcomes for the 7 participating schools indicated numerous significant differences suggesting that different assessment standards exist in one or more of the schools.

On job Performance Evaluation Reports completed before the five week training program correlated significantly with PEER ratings, Instructor ratings and final platoon standards.

The study concluded with a list of four recommendations related to future research and training.

INTRODUCTION

BACKGROUND

Prior to the integration and unification of the Army, Navy and Air Force into the Canadian Forces, each individual service conducted separate Junior and Senior non-commissioned officer (NCO) courses. Each employed different assessment methods, and the candidate's performance was reported in a variety of ways.

In 1974, one common Junior Leaders course was designed. The content of this new integrated course was derived mainly from reference material used by the individual services (1,2,3,4,5)*

Basically the aim of the Junior Leader Course is "to provide a situation of physical and mental stress wherein candidates will develop thought processes that will enhance their leadership capabilities" (8), and this is achieved by providing:

a. instruction in those military skills and knowledge (MSK) common to all junior leaders,

b. opportunities to learn and practice leadership, and

c. guidance in the development of leadership potential.

The assessment procedures currently used by the Canadian Forces Junior Leader School (CFJLS) were designed in 1974 by the attendees of the Warrant Officers academy using a modified version of the procedures employed in assessing candidates on the Basic Officers Training Course (BOTC) that is conducted at CFB Chilliwack.[1]

---

* Numbers in parenthesis refer to references.

[1] A detailed description of these assessment procedures can be found in CFP 131(3) and the CFJLS Assessment Procedures dated August 1, 1975.

Six Performance Objectives (POs) are assessed in distinct ways.

INSERT TABLE 1
about here

(a) Practical Leadership - (PO401) is assessed by

i) INSTRUCTORS' RATING ON ELEVEN CRITICAL REQUIREMENTS (CRs) OF LEADERSHIP.

Each course member is rated by his instructors on eleven dimensions (Table 2) that previous studies (3,4) suggest would discriminate successful from unsuccessful candidates. Each CR is scaled into five descriptive levels and ratings are made on the basis of a number of Critical Incident Reports that are collected and recorded throughout the training period. These reports describe in detail a given incident and the reasons for the candidates' effectiveness or ineffectiveness. They are synthesised at the end of training, converted into a rating scale score for each requirement, and summed to obtain a total CR score - worth 85% of the member's final leadership assessment grade.

INSERT TABLE 2
about here

---

ii) PEER RATINGS ON A TEN POINT LEADERSHIP SCALE

PEER rating scores are obtained as follows. First, each section member (approximately 20 men per section) ranks every other member on overall leadership ability. Then, using five descriptions of leadership effectiveness arranged into ten levels, each member assigns a rating (1 to 10) to every other member while respecting the ranking order.
PEER rating scores are worth 10% of the final leadership assessment grade.
The remaining 5% is based on a Leadership Theory examination.

(b) The Other Five Performance Objectives,

POs 402 - Aid to civil power and base defence force operations
403 - Drill
404 - Weapons training
405 - Techniques of Instruction
406 - Military Administration

fall under the rubric of Military Skills and Knowledge (MSK) and

are evaluated by written and practical examinations.

On October 10, 1975 the Canadian Forces Training System (CFTS) approached the Directorate of Personnel Applied Research (DPAR) to solicit the help of the Canadian Forces Personnel Applied Research Unit (CFPARU) in a major evaluation of the current Junior Leadership Assessment procedures.

CFPARU was explicitly requested to investigate the following:

a. Are the current 11 Critical Requirements (CRs) of leadership the best ones on which junior leaders should be assessed?

b. Are peer ratings in this context a valid form of leadership assessment?

c. Are Performance Objectives assessments standardized across schools?

d. How is previous on-job performance related to Junior Leader Training results?

Answers to these questions were sought because, effective 1977 "successful completion of this course will be mandatory for appointment to MCpl and an entrance prerequisite for the Senior Leaders Course"(6). Examination of the current Junior Leaders assessment procedures was therefore necessary to ensure that the leadership assessment procedures to be used in Junior Leaders training commencing 1977 were reliable and valid.

METHOD

Two questionnaires designed to determine the most observable, relevant and important leadership assessment requirements for Junior Leader training were completed by staff and students from 7 of the 10 Canadian Forces Junior Leader Schools. Table 3 summarizes the number of staff and students participating in the study and the information contributed by each school.

INSERT TABLE 3 ABOUT HERE

The first questionnaire,

Evaluation Of Assessment System - Junior Leaders Course. (Appendix A)

was used to reassess the original list of 25 Critical Requirements (CR) of NCO Training (4) to determine their applicability to the new unified training program.

The second questionnaire,

The Applicability Of PER Performance Requirements to Junior Leader Assessment, (Appendix B).

was designed to determine the relevance of the 17 Performance Requirements (PRs) contained on the current men's Performance Evaluation Report (PER) to assessment of Leadership at the junior leaders school.

Both staff and students were required to do three distinct ratings on the 17 Performance requirements and the 25 Critical Requirements contained in each questionnaire respectively:

1. Assess the extent to which it is possible to observe the student's behaviour for each CR or PR (PART A).
2. Rank the CRs and PRs according to the importance of each to the training program (PART B). and
3. Indicate which CRs or PRs should be deleted from the list and for what reason (PART C.)

The Standards wing located at the CFB Borden Junior Leader School administered the questionnaire and forwarded then along with the various performance objective information to the Canadian Forces Personnel Applied Research Unit for analyses.

RESULTS

A. STUDY QUESTIONNAIRES' FINDINGS

1. Data analyses of the ratings of the study questionnaires contained in Appendices A and B were utilized to answer the question: Are the 11 Critical Requirements of leadership the instructors assess course participants upon, indeed the best/most important ones. A distillation of these findings resulted in Table 2A, a new recommended list of 11 CRs on which to assess junior leaders.

INSERT TABLE 2A ABOUT HERE

---

2. Defining the CRs of leadership. During the consultations with the assessment staff at CFB Borden, it became evident that the definitions of the CRs and their corresponding behaviours varied from instructor to instructor. Current junior leader course assessment procedures don't define the CRs being assessed. This is not to say that these variations were so great as to make the assessment procedures totally unreliable. However, the extent to which these variations in definitions were left unattended, the reliability of the assessment system would be reduced. Appendix C contains definitions of each of the 11 CRs contained in Table 2A, and Appendix D lists the leadership performance rating scale levels for the same CRs.

3. The following paragraphs briefly summarize the highlights of each Study Questionnaire analysis.

B. ASSESSMENT OF 25 CRITICAL REQUIREMENTS (CR)

1. Observability in The Training Situation - (Tables A1 & A2)

The following CRs received an average observability score of less than three by either staff and/or students indicating that they were difficult to observe in the Junior Leader training context:

a. CR3 - Ensuring assistance and fair treatment of subordinates

b. CR6 - Knowing the job

c. CR16 - Ability to inspire others

d. CR18 - Telling the truth

2. Relative Importance Of The 25 CRs In Deciding Who Is The Most Effective Leader (Tables A3 and A4).

The rank order of these CRs by staff and students is very similar, (r_s=· 88, P ·001). Such high agreement reflects the perceived relevance of the CRs by both groups in the given training context; however, such rankings do not indicate the validity or reliability of assessments utilizing these CRs. The following five CRs were rated consistently high in importance by both staff and students:

a. CR1 - Assuming responsibility

b. CR5 - Instructing and briefing

c. CR7 - Correcting applying his knowledge

d. CR8 - Ensuring supervision and completion of tasks

e. CR11 - Preparation and planning for the task at hand

Those CRs rated as least important are:

a. CR25 - Admitting mistakes

b. CR18 - Telling the truth

c. CR24 - Making suggestions and giving advice

d. CR21 - Maintaining a high standard of appearance and hygiene

e. CR17 - Accepting constructive criticism

It is interesting to note that the CR assuming responsibility was ranked first by students and second by staff in importance, yet is not one of the 11 CRs that instructors currently use in assessing the leadership ability of the Canadian Forces Junior Leaders. The CR telling the truth, meaning "He is honest in what he says, and does" - was ranked on the average 25 by staff and 23 by course candidates. It appears that both staff and course members believe that whether a person tells the truth or not is of little importance in deciding who is an effective leader! It may also indicate that since it is difficult to observe, raters believed it not to be that important.

3. Relevance of the CRs to Leadership Assessment (Table A5)

Part C of both Questionnaires asked participants to indicate which CRs and PRs should be deleted from the list, and for which of the following reasons:

a. Not appropriate (valid) for junior leader assessment

b. Not observable in the training program

c. Not practical in the present form

d. Duplication of other CRs

Of the 507 staff and student study participants, only 367 indicated one or more of the 25 CRs should be deleted for the stated reasons. The average number of CRs recommended for deletion by these 367 individuals was 4.7. The reliability of the results for completion of this portion of the questionnaire are suspect due to the fact that most people did not have enough time to devote to this task. Only 50 minutes was devoted to completing Parts A, B and C of each questionnaire; it is expected that many people ran short of time on

Part C. The central finding contained in Table A5 is that many of the CRs duplicate one another.

C. ASSESSMENT OF THE 17 PERFORMANCE REQUIREMENTS

1. Observability Of The 17 PRs In The Training Situation (Tables B1 & B2)

None of the PRs received an average observability score of less than three suggesting that they could be all observed in the training environment.

2. Relative Importance of the 17 PRs in deciding who is the most effective leader (Tables B3 and B4).

The degree of agreement between staff and students rank orders of importance on the 17 PRs (r_s=.69,$p<.01$) is less than on the 25 CRs. This is probably due to the fact that the definitions of the 17 PRs are more job specific (ie., they talk about relationships between superiors and subordinates) than the training based 25 CRs, and thus create confusion in the assessor's mind concerning the appropriateness of the PR for junior leader assessment. The five PRs rated as most relevant when the staff/studen rank orders are averaged together are:

a. PR1 - Preparation and Planning

b. PR5 - Command and Self Assertion

c. PR4 - Cooperation

d. PR11 - Initiative

e. PR3 - Performance under Stress/Pressure

Those rated as least important are:

a. PR12 - Appearance and Bearing

b. PR16 - Conduct

c. PR6 - Support of Subordinates

d. PR14 - Ensuring Understanding of Assignments

e. PR8 - Knowledge of Trade/Knowledge of job when employed out of trade.

Perusal of each of the PRs histograms uncovered one major anomaly, ie., PR8 - "Knowledge of trade/knowledge of job when employed out of trade, displayed knowledge appropriate for his/her rank, trade level and experience, reasonably current, etc.," Its distribution was bi-modal. Half the raters viewed it to be very important and the other half viewed it as being relatively unimportant. It appears that if an individual perceived the item from a knowledge of trade frame of reference, he rated it as unimportant; if it was viewed from a knowledge of job when employed out of trade perspective, it was rated as very important.

3. Relevance of the PRs to leadership assessment (Table B5).

Out of 584 staff and student study participants, only 373 indicated one or more of the 17 PRs should be deleted for the reasons outlined in Part C of the Study Questionnaire. The average number of PRs recommended for deletion was 2.1 for each of the 373 people who completed Part C. As expected, Table B5 points out that six of the PRs are not practical in their present form to assess leadership in the junior leader school context.

D. THE VALIDITY OF PEER RATINGS IN THE JUNIOR LEADER TRAINING CONTEXT.

1. "Criterion-related validity is demonstrated by comparing the test scores with one or more external variables considered to provide a direct measure of the characteristic or behaviour in question".(10) This comparison usually takes the form of a correlation relating the test score (peer rating results) to a criterion measure (instructors' ratings on the 11 CRs of leadership). Validation, in the CFP 9000(2) sense, attempts to find out if the Performance Objectives are valid (11) using a questionnaire survey procedure. Both validation procedures were used in this study. The validation outcomes are displayed in Tables 4 and 5.

2. Table 4 indicates that 59% of the junior leader school staff believe that

---

Insert Table 4 about here

---

peer ratings are a valid form of leadership assessment while 51% of the course participants do not. Table 4 suggests that the students do not believe the procedure is a valid one.

3. However, comparison of the "yes" and "no" responses of the staff and students, Table 5, reveals disagreement in direction of staff and student responses for the Borden School only. In the other schools students and staff agreement on the validity of peer ratings are fairly consistent.

Insert Table 5 about here

When Peer leadership rating and Instructor leadership assessments were compared at each school statistically significant correlations ($p<.01$) were obtained (Table 6) indicating a relatively consistent agreement between staff and students on the basis used for assessment.

Insert Table 6 about here

E. STANDARDIZATION OF ASSESSMENT PROCEDURES

1. There are a total of 10 junior leadership schools established throughout Canada and Europe. The Junior Leader School at CFB Borden is tasked as standards coordinator to ensure that course materials, training standards, quality of instruction, etc., are the same across all schools.

2. This task appears an immense, if not an impossible one, given the geographic location of the schools and the fact that some schools have permanent staff and others do not.

3. To examine the degree of standardization between schools, analysis of variance and Scheffé multiple comparisons of means (13) were used to explore the following hypothesis:

a. The mean scores on the following variables for the seven participating schools do not differ significantly from one another on

(1) Total CRs scores

(2) Total leadership scores

(3) Drill grades

(4) Technique of instruction assessment

(5) Personnel administration

4. Of the 21 possible comparisons of mean total CR scores across the 7 participating schools or groups, (Tables 7 and 7a), 11 significant differences were found. For example, the mean Critical Requirements (CRs) scores of

---

INSERT TABLE 7 ABOUT HERE

---

leadership for the Halifax school ($\bar{x}$ = 69.66) differed significantly from each of the following groups:

a. Borden ($\bar{x}$ = 61.20) - $p<.01$

b. Penhold ($\bar{x}$ = 63.11) - $p<.01$

c. Esquimalt ($\bar{x}$ = 62.98) - $p<.01$

d. Europe ($\bar{x}$ = 62.60) - $p<.01$

e. Gagetown ($\bar{x}$ = 65.26) - $p<.05$

Assuming that leadership ability is evenly distributed across schools, these results suggest that the junior leader staff in Halifax assess candidates' leadership ability in a different manner than 5 out of 6 of the remaining

schools or vice versa, suggesting that uniformity in assessment across all schools is not being achieved.

5. The course training standard assigns the following weights to the performance objectives: PO401 - Leadership - 50%, PO402 - Aid to Civil power and base defense operations - 10%, PO403 - Drill - 10%, PO404 - Weapons Training - 0%, PO405 - Technique of Instruction - 10%, PO406 - Personnel Administration - 20%. Stepwise Multiple Regression was used to compare how closely schools adhered to this formula (Table 8).

INSERT TABLE 8 ABOUT HERE

From Table 8 only CFB Borden appears to be assessing individuals according to the weights specified for the respective POs. However, even at Borden there were inconsistencies across platoons in the emphasis placed on the various POs. Table 9 shows the mean performance objective results by platoon.

INSERT TABLE 9 ABOUT HERE

It is evident that standardization is not being achieved. In discussions with the assessment staff at CFB Borden it was apparent that the definitions of the CRs and their corresponding behaviours varied from instructor to instructor and that the variations in definitions reduce the reliability of the assessment system.

F. RELATIONSHIP OF PREVIOUS JOB PERFORMANCE TO JUNIOR LEADER TRAINING.

1. Peer rating, platoon standing, and instructor rating results were also significantly correlated with an external criteria, ie., total performance evaluation report (PERs) scores.

2. From Table 10 there appears to be a modest relationship between on job performance and training results.

INSERT TABLE 10 ABOUT HERE

DISCUSSIONS AND RECOMMENDATIONS

1. Integration and unification of the Army, Navy and Air Force has resulted in a streamlining of leadership training procedures. Today, there are approximately 80,000 Canadian Forces servicemen wearing the new green uniform. In 1966, prior to integration and unification, there was an 18,439 man Navy, a 43,914 man Army and 45,114 man Air Force. Even given all the changes that have occurred since 1966, the question that still needs to be answered is: Should basic officers leadership training and junior leadership training be element specific, or should everyone in the Canadian Forces meet a common standard of basic leadership skills and general service knowledge?

2. The official National Defence Headquarters policy is Green; everyone in the Canadian Forces must achieve a common standard of basic leadership skills regardless of element affiliation. With this assumption in mind, the study concluded with four recommendations:-

a. That the 11 CRs contained in Table 2A along with the contents of Appendices C and D be used by course instructors to evaluate a

candidate's leadership ability rather than the current CRs (Table 2) now being assessed.

b. Recognizing that the reliability and validity of instructors' ratings of candidates' leadership ability in the junior leader context may be suspect, a decision on the reliability and validity of Peer Ratings should be delayed until further research in this area can be carried out.

c. That consideration be given to reduce the number of junior leader schools from ten to one. This or some other alternative must be considered to assist in simplifying and standardizing the application of common assessment procedures.

d. That further study and research of the revised junior leadership assessment procedures be undertaken by a Standards Wing Officer and a military psychologist to determine:-

i. the impact of the changes implemented, and

ii. the relationship of course training results to on-job performance as measured by annual Performance Evaluation Reports (PER's) by trade or trade family.

3. Two critical requirements, assuming responsibility and command and self-assertion, were viewed by staff and students alike as being essential to assessing leadership effectiveness in the Junior Leaders School content. These were not being assessed in the leadership setting studied. Also, no operational definition of the CRs were in use. The exclusion of the two CRs just cited undoubtedly affected the validity of instructor assessments on leadership. The absence of operational definitions for each of the CRs reflects on inter-rater reliability across and within schools.

4. Given the reservations concerning instructor leadership assessments just cited, the validity of peer ratings question studied from criteria accessible in the training programme is questionnable. However, a validity check using a

Performance Evaluation Report scores determined on the job prior to attendance on the five-week Junior Leaders Course suggested that both peer ratings and instructor assessments of leadership ability have the potential to predict on-job performance. These relationships should be explored more rigorously and in greater depth in future leadership studies. Since career progression and promotion is trade or trade family related, future analyses in this regard cannot exclude this factor as was the case in the current study.

5. Standardization of assessment procedures across the seven schools studied is a real problem. This finding viewed in light of a centralized career management system has to be addressed, otherwise there can be no guarantees that the best qualified person will get the promotion or qualify for the Senior Leaders Course. Many of the findings outlined above reinforce the impression that there is a considerable ambivalence in the Canadian Forces today concerning the leadership philosophy implications of unification and integration. Until all the mixed feelings concerning "green" and environmental specific training philosophies are fully addressed, it is next to impossible to meaningfully research career management implications of leadership courses.

REFERENCES

1. Leadership: Junior Leaders Manual [CFP131(1)]. Ottawa, Ontario, Canada: National Defence Headquarters, July 1973.

2. Leadership for the Professional Officer [CFP131(2)], Ottawa, Ontario, Canada: National Defence Headquarters, October, 1966.

3. Leadership: Basic Officer Training Assessment Procedures [CFP131(3)], Ottawa, Ontario, Canada: National Defence Headquarters, January, 1976.

4. Otke, P.G., McCormack, M.D. A Study of Assessment Procedures for NCO Training (Tech. Rep. 66-6). Toronto 12, Ontario: Canadian Forces Personnel Applied Research Unit, 1107 Avenue Road, August, 1966.

5. Canadian Army Instructor's Handbook of Non-Commissioned Officer Assessment Procedures (Supplement to Tech Rep. 66-6). Toronto 12, Canadian Forces Personnel Applied Research Unit, 1107 Avenue Road, August, 1966.

6. Canadian Forces Junior Leader School Concept of Training. Trenton, Ontario: Canadian Forces Training Systems Headquarters, August, 1975.

7. Canadian Forces Junior Leader School Assessment Procedures. Trenton, Ontario: Canadian Forces Training Systems Headquarters, August, 1975.

8. Canadian Forces Junior Leader School Course Training Standard. Trenton, Ontario: Canadian Forces Training Systems Headquarters, June 1975.

9. Terry, C.M., & Otke, P.G. A Study of Assessment Procedures for Pre-Commission Training. Toronto 12, Ontario: Canadian Forces Personnel Applied Research Unit, 1964.

10. Standards for Educational and Psychological Tests and Manuals. Washington, D.C.: American Psychological Association, 1966.

11. Canadian Forces Training System - Course Training. [CFP9000(2)] Trenton, Ontario: Training Systems Headquarters, August, 1973.

12. Mallette, R. The Revision of the Basic Officer Course Assessment Procedures. (Unpublished Tech. Rep.) Toronto 12, Ontario, CFPARU, 1973.

13. Winer, B.J. Statistical Principles in Experimental Design (2nd ed.). Toronto, Ontario: McGraw Hill, 1971.

14. Hays, W.L., and Winkler, R.L. Statistics: Probability, Inference and Decision. Toronto, Ontario: Holt, Rinehard and Winston, Inc. 1971.

TABLES

LIST OF TABLES

TABLE 1 - PERFORMANCE OBJECTIVES (PO) ASSESSED DURING JUNIOR LEADERSHIP TRAINING.

TABLE 2 - CRITICAL REQUIREMENTS.

TABLE 2A - RECOMMENDED CRITICAL REQUIREMENTS.

TABLE 3 - PARTICIPATING SCHOOLS AND INFORMATION CONTRIBUTED BY EACH SCHOOL.

TABLE 4 - FACE VALIDITY OF PEER RATINGS.

TABLE 5 - FREQUENCY OF RESPONSE OF STAFF AND STUDENTS WHO ANSWERED THE QUESTION: "ARE PEER RATINGS AN APPROPRIATE (VALID) FORM OF LEADERSHIP ASSESSMENT AS USED ON THE JUNIOR LEADER COURSE?

TABLE 6 - CORRELATION OF PEER LEADERSHIP RATING WITH INSTRUCTOR LEADERSHIP ASSESSMENTS AT EACH JUNIOR LEADER SCHOOL.

TABLE 7 - MEAN PERFORMANCE OBJECTIVE SCORES BY JUNIOR LEADERSHIP SCHOOL.

TABLE 7A - ANALYSIS OF VARIANCE AND SIGNIFICANCE TESTS (SCHEFFE) COMPARING THE MEAN TOTAL CR SCORES FOR 7 JUNIOR LEADERS SCHOOLS.

TABLE 8 - STEPWISE REGRESSION BY SCHOOL USING PLATOON STANDING AS THE CRITERION AND SELECTED POs AS PREDICTORS.

TABLE 9 - COMPARISON OF MEAN PERFORMANCE OBJECTIVE (PO) SCORES BY PLATOON AT CFB BORDEN.

TABLE 10 - CORRELATION OF PERFORMANCE EVALUATION REPORT RESULTS DETERMINED PRIOR TO JUNIOR LEADER TRAINING, WITH PEER RATINGS, INSTRUCTOR RATINGS, AND PLATOON STANDING ON JUNIOR LEADER COURSE.

TABLE 1

PERFORMANCE OBJECTIVES (PO) ASSESSED DURING JUNIOR LEADERSHIP TRAINING[1]

1. PO 401 - Practical Leadership
2. PO 402 - Aid to civil power and base defence force operations
3. PO 403 - Drill
4. PO 404 - Weapons Training
5. PO 405 - Techniques of Instruction
6. PO 406 - Military Administration

[1]For this study data was only collected on POs 401,403,405 and 406.

TABLE 2

CRITICAL REQUIREMENTS[a]

1. Preparation & Planning
2. Briefing Others
3. Adaptability
4. Supervision
5. Job Knowledge
6. Initiative
7. Stress - Mental & Physical
8. Providing & Seeking Advice
9. Self-Expression
10. Co-operation
11. Accepting Criticism

---

[a]These 11 CRs were being used prior to this study. No operational definitions for each CR existed.

TABLE 2A

RECOMMENDED CRITICAL REQUIREMENTS

DOES HE/SHE

1. Assume Responsibility?
2. Take Command and Assert Him/Her Self?
3. Prepare and Plan?
4. Brief Others?
5. Supervise?
6. Display Adaptability/Initiative?
7. Communicate Effectively?
8. Perform Effectively Under Stress?
9. Correctly Apply Knowledge?
10. Cooperate with Others?
11. Learn from Experience?

TABLE 3

PARTICIPATING SCHOOLS AND INFORMATION CONTRIBUTED BY EACH SCHOOL.[a]

School	Course Number	Performance Objective Data	25 CRs Questionnaire		17 PRs Questionnaire	
			Staff	Student	Staff	Student
1. CFB Borden	7506 7601	373	32	186	31	208
2. CFB Penhold	7506 7601	244	16	147	17	150
3. CFB Esquimalt	7506 7507	63	5	47	5	48
4. CFB Europe[d]	7504 7601	124	13	30	12	30
5. CFB Halifax[c,d]	7506	128	—	31	5	26
6. CFB Summerside[c]	7506	44	—	—	9	44
7. CFB Gagetown[b]	7502	42	—	—	—	—
Total Number of People		1018	66	441	79	506

a. The staffs of each of the participating schools were asked to complete both Study Questionnaires. Half the students on each course were required to complete the 25 CRs Questionnaire, the other half the 17 PRs Questionnaire.

b. CFB Gagetown did not complete the two Study Questionnaires. They were not asked to do so.

c. CFBs Summerside and Halifax staff and students completed only one of the two Study Questionnaires.

d. Two courses of performance objective data were analyzed for CFB Europe and CFB Halifax, yet only one course of students at each school contributed study questionnaire

TABLE 4

FACE VALIDITY OF PEER RATINGS[1]

School	Staff Answers			Student Answers		
	Yes	No	Not Answered	Yes	No	Not Answered
Borden	18	12	2	54	119	13
Penhold	5	9	2	40	60	47
Esquimalt	5	0	0	24	12	11
Europe (Lahr)	11	1	1	21	7	2
Halifax				5	26	0
Totals	39	22	5	144	224	73
	(59)[2]	(33)	(9)	(33)	(51)	(17)

[1]The numbers that appear in this table refer to the frequency of answers given to the following question:

a. Are PEER RATINGS an appropriate (valid) form of leadership assessment as used on the junior leader course?

[2]The numbers contained in parentheses are percentages. They do not total 100% due to rounding errors.

1349

TABLE 5

FREQUENCY OF RESPONSE OF STAFF AND STUDENTS WHO ANSWERED THE QUESTION: "ARE PEER RATINGS AN APPROPRIATE (VALID) FORM OF LEADERSHIP ASSESSMENT AS USED ON THE JUNIOR LEADER COURSE?"

	SCHOOL		YES	NO	
1.	BORDEN	STAFF	18	12	$X^2 = 8.041$ $p < .005$
		STUDENTS	54	119	df=1 (see Note 1 below)
			YES	NO	
2.	PENHOLD	STAFF	5	9	$X^2 = .0002$ N.S.
		STUDENTS	40	60	df=1
			YES	NO	
3.	ESQUIMALT	STAFF	5	0	$X^2 = 1.0212$ N.S.
		STUDENTS	24	12	df=1
			YES	NO	
4.	LAHR	STAFF	11	1	$X^2 = 0.6027$ N.S.
		STUDENTS	21	7	df=1
			YES	NO	
5.	ABOVE FOUR SCHOOLS COMBINED	STAFF	39	22	$X^2 = 12.167$ $p < .001$
		STUDENTS	144	224	df=1

Note 1. All X^2 (CHI-SQUARED) values corrected for continuity using Yates' correction. (Hays and Winkler, 1971 p797)

TABLE 6

CORRELATION OF PEER LEADERSHIP RATING WITH INSTRUCTOR LEADERSHIP ASSESSMENTS AT EACH JUNIOR LEADER SCHOOL

School	N	r
CFB Penhold	244	.38 *
CFB Gagetown	42	.46 *
CFB Halifax	149	.47 *
CFB Borden	373	.47 *
CFB Summerside	44	.50 *
CFB Esquimalt	63	.59 *
CFB Lahr	103	.60 *

$p < .01$

1351

TABLE 7

MEAN PERFORMANCE OBJECTIVE SCORES BY JUNIOR LEADERSHIP SCHOOL[1]

POs \ Schools	Borden	Penhold	Esquimalt	Europe	Halifax	Summerside	Gagetown	Grand Mean
1. Critical Requirements of Leadership	61	63	63	62	70	67	65	63
2. Personnel Administration	79	81	89	75	80	80	79	80
3. Techniques of Instruction	71	64	67	73	67	76	77	69
4. Drill	68	67	75	73	77	70	68	70

[1]Means rounded to nearest whole number. Those underlined indicate a significant difference ($p < .01$) between schools on that respective variable.

1225

ANALYSIS OF VARIANCE AND SIGNIFICANCE TESTS (SCHEFFE) COMPARING THE MEAN TOTAL CR SCORES FOR 7 JUNIOR LEADERS SCHOOLS

GROUP		[illegible]	[illegible]	VARIANCE	S.D.
1	Borden	272.	61.[illegible]	[illegible]	6.6333
2	Penhold	244.	63.1139	3[illegible]	4.2499
3	Esquimalt	63.	62.3841	33.5665	5.7937
4	Europe	124.	62.3968	31.3994	5.6133
5	Halifax	136.	69.6838	51.3771	7.1673
6	Summerside	44.	57.4545	26.4264	5.1414
7	Gagetown	42.	65.2619	18.7363	4.3285
TOTAL		1118.	63.4421	47.3738	6.8829

HOMOGENITY OF VARIANCE TEST CHI[illegible]= 21.9853 PROBABILITY= 0.0012

ANALYSIS OF VARIANCE

SOURCE	SS	MS	DF	F	P
GROUPS	[illegible]	1298.17	6.	32.46	0.000000
ERROR	[illegible]	40.[illegible]	1[illegible].		

PROBABILITY MATRIX FOR SCHEFFE MULTIPLE COMPARISON OF MEANS

	1	2	3	4	5	6	7
1	1.0000	0.7358*	0.6382	0.6051	0.0000**	0.0000**	0.0169*
2	0.7358	1.0000	1.0000	0.9969	0.0000**	0.0179*	0.6611
3	0.6382	1.0000	1.0000	0.9969	0.0000**	0.0449*	0.7741
4	0.6051	0.9969	0.9999	1.0000	0.0000**	0.0041**	0.4733
5	0.0000	0.0000	0.0000	0.0000	1.0000	0.6839	0.0189*
6	0.0000	0.0179	0.0449	0.0041	0.6839	1.0000	0.8588
7	0.0169	0.6611	0.7741	0.4733	0.0189	0.8588	1.0000

1226

1354

*p <.05

**p <.01

1355

TABLE 8

STEPWISE REGRESSION BY SCHOOL USING PLATOON STANDING AS THE CRITERION AND SELECTED POs AS PREDICTORS[1]

SCHOOL	N	Predictors in order of contribution to multiple R and % of increase of R^2					Total R^2
Borden	373	CR	PA	PR	DRILL	T of I	
		(43)	(17)	(11)	(0)	(1)	.72
Penhold	244	CR	DRILL	PA	T of I	PR	
		(44)	(10)	(6)	(1)	(2)	.63
Esquimalt	63	CR	DRILL	PA	PR	T of I	
		(56)	(19)	(3)	(1)	(0)	.79
Lahr	124	T of I	CR	PA	DRILL	PR	
		(40)	(16)	(5)	(4)	(6)	.71
Halifax	128	PA	DRILL	CR	T of I	PR	
		(27)	(12)	(3)	(1)	(1)	.44
Summerside	44	CR	T of I	PA	DRILL	PR	
		(52)	(21)	(16)	(4)	(0)	.93
Gagetown	42	T of I	CR	PA	DRILL	PR	
		(48)	(20)	(16)	(11)	(1)	.96

[1]Numbers in parenthesis indicate the percentage increase to R^2 contributed by that variable.

TABLE 9

COMPARISON OF MEAN PERFORMANCE OBJECTIVE (PO) SCORES BY PLATOON AT CFB BORDEN[1]

Performance Objectives	Platoon 1	2	3	4	Grand Mean
1. Critical Requirements of Leadership	63	55	63	61	61
2. Personnel Administration	73	79	78	80	77
3. Techniques of Instruction	78	70	72	72	73
4. Drill	69	65	71	69	68

[1]Means underlined indicate a significant difference ($p<.01$) between platoons on that respective variable.

TABLE 10

CORRELATION OF PERFORMANCE EVALUATION REPORT RESULTS DETERMINED PRIOR TO JUNIOR LEADER TRAINING, WITH PEER RATINGS, INSTRUCTOR RATINGS, AND PLATOON STANDING ON JUNIOR LEADER COURSE

Group	Peer Ratings	Instructors Ratings	Platoon Standing
Privates (n=42) PERS	.26	.36*	.39**
Corporals (n=52) PERS	.37**	.32*	.46**

*p<.05

**p<.01

APPENDICES

TABLES APPEARING IN APPENDICES

TABLE A1 - MEAN RATINGS OF DEGREE OF OBSERVABILITY BY STAFF AND STUDENTS OF THE 25 CRITICAL REQUIREMENTS.

TABLE A2 - STAFF AND STUDENTS RANK ORDER OF DEGREE OF OBSERVABILITY OF THE 25 CRITICAL REQUIREMENTS DURING THE COURSE.

TABLE A3 - MEAN RATINGS OF RELATIVE IMPORTANCE OF THE 25 CRITICAL REQUIREMENTS TO LEADERSHIP EFFECTIVENESS.

TABLE A4 - STAFF AND STUDENTS RANK ORDER OF RELATIVE IMPORTANCE OF THE 25 CRITICAL REQUIREMENTS TO LEADERSHIP EFFECTIVENESS.

TABLE A5 - RELEVANCE OF THE CRs TO LEADERSHIP ASSESSMENT FOR STAFF AND STUDENTS.

TABLE A6 - FREQUENCY COUNT OF EXTENT OF OBSERVATION OF CRs FOR STAFF.

TABLE B1 - MEAN RATINGS OF DEGREE OF OBSERVABILITY BY STAFF AND STUDENTS OF THE 17 PERFORMANCE REQUIREMENTS (PR) OF LEADERSHIP.

TABLE B2 - RANK ORDER BY STAFF AND STUDENTS OF THE DEGREE OF OBSERVABILITY OF THE 17 PERFORMANCE REQUIREMENTS (PR) OF LEADERSHIP.

TABLE B3 - MEAN RATINGS OF RELATIVE IMPORTANCE OF THE 17 PERFORMANCE REQUIREMENTS (PR) TO LEADERSHIP EFFECTIVENESS.

TABLE B4 - STAFF AND STUDENTS RANK ORDER OF THE RELATIVE IMPORTANCE OF THE 17 PERFORMANCE REQUIREMENTS TO LEADERSHIP EFFECTIVENESS.

TABLE B5 - RELEVANCE OF PRs TO LEADERSHIP ASSESSMENT FOR STAFF AND STUDENTS.

TABLE B6 - FREQUENCY COUNT OF EXTENT OF OBSERVATION OF PRs STAFF (N=75) AND STUDENTS (N=480).

Appendix A

QUESTIONNAIRE USED TO RE-ASSESS THE RELEVANCE OF THE ORIGINAL 25 CRITICAL REQUIREMENTS (CR) TO THE PRESENT JUNIOR LEADERS TRAINING PROGRAM

EVALUATION OF ASSESSMENT SYSTEM - JUNIOR LEADERS COURSE

Introduction

1. The last major research study of NCO leadership assessment procedures was completed in August 1966. The Junior Leaders Course (JLC) today differs from the program given in previous years. Consequently, non-researched adjustments to the assessment procedures were made; ie., we now assess 11 Critical Requirements (CRs) of leadership instead of twenty-five. The purpose of this study is to reassess the original 25 CRs to determine their relevance to the present training program, and to ensure that the 11 CRs now being used are indeed the best ones to discriminate the successful from the unsuccessful junior leader.

2. How can you help? Because of your extensive participation in the JLC program (as instructor or student) you can best assess the observability, importance and relevance of the CRs used in the assessment system.

Instructions

3. On pages 5&6 is a list of the original 25 CRs used in past junior and senior NCO leadership assessment procedures. Pages three and four contain operational definitions of each CR; these definitions should assist you in your ratings of these leadership assessment requirements. You are requested to make three distinct ratings on the 25 CRs, and record your replies in the A, B and C portions of the rating scale appearing on pages 5 and 6.

PART A - Ideally, the training program should have provided opportunities for you to observe directly students' performance under these requirements. Indicate under the A rating section on page 5 to WHAT EXTENT you would be able to directly observe these CRs in your students or peers during this training. Circle the appropriate number showing this for each of the requirements.

PART B - Here we are interested in finding out the relative importance to you of each of these critical requirements in deciding who is an effective leader. Study the list of the 25 CRs carefully. Four steps are involved in this rating process.

STEP ONE - Look at the 25 critical requirements and select the one you think is the MOST important for this training.

STEP TWO - Insert the number 1. beside this critical requirement in the PART B column, and having done this draw a line through this critical requirement.

STEP THREE - From the remaining list, select the one you think is the LEAST important for this training. Insert the number 25 beside this critical requirement in the PART B column, and having done this draw a line through this critical requirement.

STEP FOUR - From the remaining list, select the CR you think is the NEXT MOST important and mark a 2. beside it. Continue in the same manner for the NEXT LEAST important CR and mark a 24 beside it. Continue doing this until all CRs have been evaluated in this manner, ie. 3 and 23, 4 and 22, 5 and 21, 6 and 20, 7 and 19, 8 and 18, 9 and 17, 10 and 16, 11 and 15, 12 and 14, 13.

PART C - LASTLY - Read the list of CRs and select one you strongly feel should be removed from the assessment system. Select below one number indicating the reason for your decision. Insert this number in Column C opposite the CR number shown on the right of this column. Repeat this procedure for any other CRs you strongly feel should be deleted.

1. Not appropriate (valid) for Junior Leadership Assessment.
2. Not observable within the training program.
3. Not practical in its present form.
4. Duplication of other CRs.

1. Assuming Responsibility

- Takes upon himself to do assigned tasks without complaining
- He acts and accepts the consequences of his actions
- Fulfills his duties consistently

2. Taking action on his own

- He anticipates requirements, and is self-propelled in fulfilling them.
- Tries novel ideas to achieve results
- Steps in and takes action when the need arises, even in face of opposition

3. Ensuring assistance and fair treatment of subordinates

- He facilitates cooperation by sharing his knowledge and appreciating the feelings of others. He is considerate
- Demands no more than he himself is prepared to give
- Ensures equality of treatment and applies discipline when needed

4. Making an effort to cooperate with others, and working as a member of a team

- He cooperates with associate who is in charge
- Provides assistance to others whenever he can without trying to take over
- Puts the needs of others ahead of his own

5. Instructing and Briefing

- He provides others with the needed information to do the job
- Gives clear and concise instructions demonstrating where appropriate

6. Knowing the job

- He has in-depth knowledge of what is required of him as a Junior leader when he is completing a task.

7. Correctly applying his knowledge

- Puts ideas into practice
- Demonstrates his ability to utilize new knowledge and experience
- Does not repeatedly make the same mistake

8. Ensuring supervision, and completion of tasks

- He makes sure the jobs for which he is responsible are properly executed by establishing priorities, assigning the task to the right man, and spot-checking to ensure task is understood and being carried out

9. Displaying proper personal conduct

- He leads by example, is self-determing, and has the ability to handle most of his personal problems on his own

10. Performing successfully under periods of emotional stress

- Maintains pursuit of goals under difficult conditions
- Achieves acceptable results even under stre

11. Preparation and planning for the task at hand

- He anticipates events/problems, and the needed resources to accomplish the task
- Plans flexibly by adjusting actions to the situation
- Subdivides the job into specific tasks and delegates appropriately.

12. Ability to adapt to new problems and situations

- He is open minded, flexible, and receptive to ideas from superiors and subordinates

13. Supporting superiors and subordinates even though he disagrees

- Can suspend his differing views, and press on with getting the job done
- Demonstrates loyalty and solidarity for his associates

14. Maintaining proper interpersonal relationships - exercising discipline and control of subordinates

- Sets high standards for himself and others leads by example and inspires his subordinates to do the same
- Applies discipline fairly, and only when needed.
- Makes an effort to amend strained personal relations in the group
- Utilizes chain of command in the disciplining and control of subordinates

Delegating tasks to others

Ensures he can attend his main functions by assigning less vital tasks to others

Ability to inspire others

Has personal qualities (charisma) that inspire others to do an above average job

Accepting constructive criticism

Shows receptiveness to constructive guidance and profits from it

Converts mistakes into a source of learning on subsequent tasks

Telling the truth

He is honest in what he says, and does

Ensuring the understanding of assigned tasks

He ensures understanding of the problem by requesting feedback, practicing the class on the task, and correcting errors

Complying with rules and orders even though he disagrees

Foregoes personal comfort or opinion to carry out superior's requests
Takes immediate action on request by superiors
Makes sacrifices (if necessary) to abide by rules and regulations

Maintaining a high standard of appearance and hygiene

His appearance and bearing contributes to his effectiveness as a serviceman

22. Performing successfully under periods of physical stress

- Retains self-control under difficult conditions
- Maintains pursuit of goals under difficult conditions
- Achieves acceptable results even under stress

23. Expressing himself clearly and concisely

- Gives clear and complete orders
- Provides others with accurate and timely information orally and in writing if necessary
- Ensures others have understood his instructions

24. Making suggestions and giving advice

- He discusses problems and works out solution with others
- Incorporates others' suggestions and advice into his own

25. Admitting mistakes

- Admits his mistakes and accepts their consequences

1365

EVALUATION OF JLC ASSESSMENT SYSTEM

CRITICAL ELEMENTS	PART A EXTENT OF DIRECT OBSERVATION					PART B	PART C
	Never	Barely	Adequate	Fairly Often	Frequently		
1. Assuming responsibility	1	2	3	4	5		1
2. Taking action on his own	1	2	3	4	5		2
3. Ensuring assistance and fair treatment of subordinates	1	2	3	4	5		3
4. Making an effort to cooperate with others and working as a member of a team	1	2	3	4	5		4
5. Instructing and briefing	1	2	3	4	5		5
6. Knowing the job	1	2	3	4	5		6
7. Correctly applying his knowledge	1	2	3	4	5		7
8. Ensuring supervision and completion of tasks	1	2	3	4	5		8
9. Displaying proper personal conduct	1	2	3	4	5		9
10. Performing successfully under periods of emotional stress	1	2	3	4	5		10
11. Preparation and planning for the task at hand	1	2	3	4	5		11
12. Ability to adapt to new problems and situations	1	2	3	4	5		12
13. Supporting superiors and associates	1	2	3	4	5		13
14. Maintaining proper interpersonal relationships - exercising discipline and control of subordinates	1	2	3	4	5		14
15. Delegating tasks to others	1	2	3	4	5		15
16. Ability to inspire others	1	2	3	4	5		16
17. Accepting constructive criticism	1	2	3	4	5		17
18. Telling the truth	1	2	3	4	5		18
19. Ensuring the undertaking of assigned tasks	1	2	3	4	5		19
20. Complying with rules and orders	1	2	3	4	5		20

intaining a high standard appearance and hygiene	1	2	3	4	5	21	
forming successfully er periods of physical ress	1	2	3	4	5	22	
pressing himself clearly d concisely	1	2	3	4	5	23	
king suggestions and ving advice	1	2	3	4	5	24	
lmitting mistakes	1	2	3	4	5	25	

A Please circle one of the following:
(1) I am a student
(2) I am a staff member

B Are PEER RATINGS an appropriate (valid) form of leadership assessment as used on the JLC? Circle your answer.
(1) Yes
(2) No

C Location of Junior Leaders' School - Circle One

1. CFB Borden
2. CFB Penhold
3. CFB Esquimalt
4. CFB Europe
5. CFB Halifax
6. CFB Summerside

D Please feel free to elaborate on your answer to the peer rating question, and any other aspect of this study that you believe needs comment. Please write your answer in the space below.

THANK YOU FOR YOUR ASSISTANCE.

TABLE A1

MEAN RATINGS OF DEGREE OF OBSERVABILITY BY STAFF AND STUDENTS OF THE 25 CRITICAL REQUIREM

Critical Requirements	STAFF (n=65)		STUDENTS (n=426	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1. Assuming responsibility	3.89	.75	3.66	.93
2. Taking action on his own	3.26	.88	3.04	.95
3. Ensuring assistance and fair treatment of subordinates	2.86	1.02	3.01	.93
4. Making an effort to cooperate with others and working as a member of a team	3.34	.64	3.23	.70
5. Instructing and briefing	4.14	.74	3.70	.89
6. Knowing the job	3.57	1.10	2.95	1.05
7. Correctly applying his knowledge	3.80	.81	3.39	.85
8. Ensuring supervision and completion of tasks	4.17	.74	3.62	.94
9. Displaying proper personal conduct	3.42	.99	3.56	.98
10. Performing successfully under periods of emotional stress	3.26	1.09	3.38	1.01
11. Preparation and planning for task at hand	4.18	.82	3.65	.90
12. Ability to adapt to new problems and situations	3.66	.88	3.60	.87
13. Supporting superiors and associates	3.17	.97	3.40	.94
14. Maintaining proper interpersonal relationships - exercising discipline and control of subordinates	3.02	1.10	3.09	1.02
15. Delegating tasks to others	3.95	.92	3.64	.93
16. Ability to inspire others	2.88	.95	2.89	.90
17. Accepting constructive criticism	3.85	.96	3.67	.90
18. Telling the truth	2.66	1.18	3.31	1.15
19. Ensuring the undertaking of assigned tasks	3.60	.97	3.50	.92
20. Complying with rules and orders	3.72	.94	3.93	.92
21. Maintaining a high standard of appearance and hygiene	4.05	1.00	4.15	.93
22. Performing successfully under periods of physical stress	3.48	1.08	3.59	.94
23. Expressing himself clearly and concisely	4.02	.77	3.66	.87
24. Making suggestions and giving advice	3.48	.90	3.39	.93
25. Admitting mistakes	3.31	1.05	3.29	.99

TABLE A2

TAFF AND STUDENTS RANK ORDER OF DEGREE OF OBSERVABILITY OF THE 25 CRITICAL REQUIREMENTS JRING THE COURSE.

Critical Requirements	Rank Order a) b) STAFF (n=65)	STUDENTS (n=426)
Assuming responsibility	7	6
Taking action on his own	20	22
Ensuring assistance and fair treatment of subordinates	24	23
Making an effort to cooperate with others and working as a member of a team	17	20
Instructing and briefing	3	3
Knowing the job	13	24
Correctly applying his knowledge	9	16
Ensuring supervision and completion of tasks	2	9
Displaying proper personal conduct	16	12
Performing successfully under periods of emotional stress	19	17
Preparation and planning for task at hand	1	7
Ability to adapt to new problems and situations	11	10
Supporting superiors and associates	21	14
Maintaining proper interpersonal relationships - exercising discipline and control of subordinates	22	21
Delegating tasks to others	6	8
Ability to inspire others	23	25
Accepting constructive criticism	8	4
Telling the truth	25	18
Ensuring the undertaking of assigned tasks	12	13
Complying with rules and orders	10	2
Maintaining a high standard of appearance and hygiene	4	1
Performing successfully under periods of physical stress	15	11
Expressing himself clearly and concisely	5	5
. Making suggestions and giving advice	14	15
. Admitting mistakes	18	19

a) Spearman's rank order correlation = .82 p<.001

b) Kendall's tau = .60 p<.0001

MEAN RATINGS OF RELATIVE IMPORTANCE OF THE 25 CRITICAL REQUIREMENTS TO LEADERSHIP EFFECTIVENESS

Critical Requirements	STAFF (n=65)		STUDENTS (n=426)	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.
1. Assuming responsibility	5.95	5.60	4.52	5.12
2. Taking action on his own	10.08	6.34	12.78	7.53
3. Ensuring assistance and fair treatment of subordinates	16.17	5.74	12.28	6.24
4. Making an effort to cooperate with others and working as a member of a team	10.73	3.87	9.86	4.07
5. Instructing and briefing	6.65	4.99	9.58	6.32
6. Knowing the job	10.14	7.34	12.40	6.34
7. Correctly applying his knowledge	8.61	5.73	10.67	6.42
8. Ensuring supervision and completion of tasks	6.86	4.45	10.25	5.91
9. Displaying proper personal conduct	16.18	5.55	15.01	6.71
10. Performing successfully under periods of emotional stress	14.38	6.50	13.33	6.85
11. Preparation and planning for task at hand	3.36	3.47	8.12	5.88
12. Ability to adapt to new problems and situations	10.80	5.29	10.74	5.85
13. Supporting superiors and associates	15.18	6.09	14.64	6.38
14. Maintaining proper interpersonal relationships - exercising discipline and control of subordinates	15.53	7.27	13.97	7.04
15. Delegating tasks to others	10.00	5.90	12.56	6.17
16. Ability to inspire others	15.05	6.20	14.71	7.41
17. Accepting constructive criticism	16.08	5.84	15.38	5.67
18. Telling the truth	20.83	4.20	17.12	7.01
19. Ensuring the undertaking of assigned tasks	12.91	5.99	13.25	5.53
20. Complying with rules and orders	13.64	5.21	15.03	6.29
21. Maintaining a high standard of appearance and hygiene	19.09	5.17	16.51	7.14
22. Performing successfully under periods of physical stress	14.14	6.01	14.60	6.75
23. Expressing himself clearly and concisely	12.29	6.35	12.96	6.20
24. Making suggestions and giving advice	17.92	5.11	17.21	5.75
25. Admitting mistakes	20.27	3.97	18.52	6.13

TABLE A4

STAFF AND STUDENTS RANK ORDER OF RELATIVE IMPORTANCE OF THE 25 CRITICAL REQUIREMENTS TO LEADERSHIP EFFECTIVENESS

Critical Requirements	Rank Order a) b) STAFF	STUDENTS
Assuming responsibility	2	1
Taking action on his own	7	11
Ensuring assistance and fair treatment of subordinates	20	8
Making an effort to cooperate with others and working as a member of a team	9	4
Instructing and briefing	3	3
Knowing the job	8	9
Correctly applying his knowledge	5	6
Ensuring supervision and completion of tasks	4	5
Displaying proper personal conduct	21	19
Performing successfully under periods of emotional stress	15	14
Preparation and planning for task at hand	1	2
Ability to adapt to new problems and situations	10	7
Supporting superiors and associates	17	17
Maintaining proper interpersonal relationships - exercising discipline and control of subordinates	18	15
Delegating tasks to others	6	10
Ability to inspire others	16	18
Accepting constructive criticism	19	21
Telling the truth	25	23
. Ensuring the undertaking of assigned tasks	12	13
. Complying with rules and orders	13	20
. Maintaining a high standard of appearance and hygiene	23	22
. Performing successfully under periods of physical stress	14	16
. Expressing himself clearly and concisely	11	12
. Making suggestions and giving advice	22	24
. Admitting mistakes	24	25

a) Spearman's rank order correlation =.88 p<.001

[illegible] = .75 p<.0001

TABLE A5

RELEVANCE OF THE CRs TO LEADERSHIP ASSESSMENT FOR STAFF AND STUDENTS (N=367)

Critical Requirements	Main Reason for Dropping[1]				Frequency of Total Ratings	
	1	2	3	4		%
1. Assuming responsibility				X	34	2.03
2. Taking action on his own				X	89	5.30
3. Ensuring assistance and fair treatment of subordinates		X			96	5.72
4. Making an effort to cooperate with others and working as a member of a team				X	26	1.55
5. Instructing and briefing				X	55	3.28
6. Knowing the job	X				120	7.15
7. Correctly applying his knowledge				X	45	2.68
8. Ensuring supervision and completion of tasks				X	45	2.68
9. Displaying proper personal conduct				X	71	4.23
10. Performing successfully under periods of emotional stress				X	138	8.22
11. Preparation and planning for task at hand				X	18	1.07
12. Ability to adapt to new problems and situations				X	27	1.61
13. Supporting superiors and associates				X	88	5.24
14. Maintaining proper interpersonal relationships - exercising discipline and control of subordinates				X	59	3.52
15. Delegating tasks to others				X	32	1.91
16. Ability to inspire others		X			103	6.14
17. Accepting constructive criticism				X	39	2.32
18. Telling the truth				X	121	7.21
19. Ensuring the undertaking of assigned tasks		X			57	3.40
20. Complying with rules and orders				X	47	2.80
21. Maintaining a high standard of appearance and hygiene				X	42	2.50
22. Performing successfully under periods of physical stress				X	113	6.73
23. Expressing himself clearly and concisely				X	49	2.92
24. Making suggestions and giving advice				X	65	3.87
25. Admitting mistakes				X	99	5.90
					N = 1678	

1. Legend

1 Not communication (valid) for H assessment 3 Not practical in its present form

TABLE A6

FREQUENCY COUNT OF EXTENT OF OBSERVATION OF CRs FOR STAFF (N=65)

Critical Requirements	Never 1	2	Adequate 3	4	Frequently 5
suming responsibility	0	2	16	34	13
king action on his own	1	12	25	23	4
suring assistance and fair treatment of bordinates	6	18	23	15	3
king an effort to cooperate with others and rking as a member of a team	8	17	25	13	2
structing and briefing	0	0	14	28	23
owing the job	2	11	14	24	14
rrectly applying his knowledge	0	2	23	26	14
suring supervision and completion of tasks	0	0	13	28	14
splaying proper personal conduct	2	10	20	25	8
erforming successfully under periods of motional stress	1	18	20	15	11
reparation and planning for the task at hand	0	2	11	25	27
bility to adapt tonew problems and situations	1	4	22	27	11
upporting superiors and associates	3	12	26	19	5
aintaining proper interpersonal relationships - xercising discipline and control of subord- nates	5	16	25	11	8
elegating tasks to others	1	3	14	27	20
bility to inspire others	4	20	23	16	2
ccepting constructive criticism	1	5	15	26	18
elling the truth	11	20	21	6	7
nsuring the undertaking of assigned tasks	2	7	16	30	10
omplying with rules and orders	1	4	22	13	15
aintaining a high standard of appearance and ygiene	1	3	16	17	28
Performing successfully under periods of physical stress	2	8	28	11	16
Expressing himself clearly and concisely	0	1	16	29	19
Making suggestions and giving advice	0	7	31	16	1
Admitting mistakes	2	14	20	20	9

Appendix B

QUESTIONNAIRE USED TO EVALUATE THE APPLICABILITY OF PERFORMANCE EVALUATION REPORT REQUIREMENTS TO JUNIOR LEADER ASSESSMENT

THE APPLICABILITY OF PER PERFORMANCE REQUIREMENTS TO JUNIOR LEADER ASSESSMENT

Introduction

1. NDHQ has tasked the Canadian Forces Personnel Applied Research Unit to evaluate all assessment systems. The Performance Evaluation Reports (PERs) and the Junior Leaders' Course (JLC) assessment procedures are just two of many assessment systems in operation in the Canadian Forces today. The purpose of this study is to determine the relevance of the 17 Performance Requirements contained on the current PER to the assessment of leadership at the junior leaders school.

2. An extensive job analysis conducted by interviewing personnel on course, on the job, and their superiors and subordinates, resulted in the identification of 25 types of behaviour essential to leadership performance. The 17 Performance Requirements (CRs) contained on the Men's PER, along with the 11 Critical Requirements (CRs) used in leadership assessment at the JLS were derived from these original 25 CRs.

3. The purpose of the Junior Leaders' Course is to provide those attending with the necessary knowledge and skills required "to function effectively in the Junior Leaders' role". On the job, this will be assessed by means of the PER. Where job and junior leader course objectives are common, the assessment systems should also be common.

Instructions

4. On page 3 appears a list of definitions of the 17 performance requirement (PRs) now used in the evaluation of level 4 privates to chief warrant officers; read these before you complete the study questionnaire. Because of your extensive participation in the Junior Leaders program (as instructor or student) you can best assess the observability (PART A), importance (PART B), and relevance (PART C) of these 17 PRs to the Junior Leaders' Training Program.

PART A - Ideally, the training program should have provided opportunities for you to observe directly students' performance under these requirements. Indicate under the A rating section on page 4 to WHAT EXTENT you would be able to directly observe these PRs in your students or peers during this training. Circle the appropriate number showing this for each of the requirements.

PART B - Here we are interested in finding out the relative importance to you of each of these performance requirements in deciding who is an effective leader and/or the most productive worker. Study the list of 17 PRs carefully. Four steps are involved in this rating process:

STEP ONE - look at the 17 performance requirements on page 4 and select the one you think is the MOST important for this training.

STEP TWO - Insert the number 1 beside this performance requirement in the PART B column, and having done this, draw a line through this performance requirement.

STEP THREE - From the remaining list, select the one you think is the LEAST important for this training. Insert the number 17 beside this performance requirement in the PART B column, and having done this, draw a line through this performance requirement.

STEP FOUR - From the remaining list, select the PR you think is the NEXT MOST important and mark a 2 beside it. Continue in the same manner for the NEXT LEAST important PR and mark a 16 beside it. Continue doing this until all PRs have been evaluated in this manner, ie., 3 and 15, 4 and 14, 5 and 13, 6 and 12, 7 and 11, 8 and 10, 9.

PART C - LASTLY - Read the list of PRs and select one you strongly feel should be removed from the assessment system. Select below one number indicating the reason for your decision. Insert this number in Column C opposite the PR number shown on the right of this column. Repeat this procedure for any other PRs you strongly feel should be deleted.

1. Not appropriate (valid) for Junior Leadership Assessment
2. Not observable within the training program
3. Not practical in its present form
4. Duplication of other PRs

DEFINITIONS OF PER PERFORMANCE REQUIREMENTS

1. Preparation and Planning. Adequately planned ahead and determined task priorities with little or no assistance.

2. Delegation. Delegated tasks appropriately to his subordinates who handled them with only a few problems.

3. Performance Under Stress/Pressure. Maintained an acceptable level of effectiveness in spite of stress/pressure.

4. Cooperation. Cooperated with peers and superiors, contributing to group effectiveness by working as member of a team.

5. Command and Self Assertion. Asserted himself/herself sufficiently to maintain adequate control of his/her juniors with little or no guidance.

6. Support of Subordinates. Advised and assisted subordinates on questions relating to trade and career development and personal welfare when brought to his/her attention.

7. Briefing Others. Usually provided sufficient information for subordinates to complete their tasks with a minimum of delay or confusion.

8. Knowledge of Trade/Knowledge of Job When Employed out of Trade. Displayed knowledge appropriate for his/her rank, trade level and experience. Reasonably current.

9. Ability to Apply His/Her Knowledge. His/her work was of an acceptable quality with few errors and with most jobs completed on time.

10. Adaptability. Met new problems and situations adequately, developing new solutions when familiar methods failed.

11. Initiative. Took necessary action without serious delay.

12. Appearance and Bearing. His/her appearance and bearing contributed to his effectiveness as a Serviceman/Servicewoman.

13. Supervision. Usually ensured that tasks assigned to subordinates were completed on time and were free from errors.

14. Ensuring Understanding of Assignments. Ensured subordinates' understanding of major points of instruction.

15. Responsibility. Accepted assigned responsibility.

16. Conduct. Complied with regulations and displayed acceptable level of personal conduct.

17. Learning From Experience. Sought and accepted advice whenever necessary, maintaining, and at times improving, his proficiency on basis of experience.

EVALUATION OF JUNIOR LEADERSHIP ASSESSMENT
FROM A PR FRAME OF REFERENCE

PERFORMANCE REQUIREMENTS	PART A EXTENT OF DIRECT OBSERVATION					PART B	PART C
	Never	Barely Adequate		Fairly Often	Frequently		
1. Preparation and Planning	1	2	3	4	5		1
2. Delegation	1	2	3	4	5		2
3. Performance Under Stress/ Pressure	1	2	3	4	5		3
4. Cooperation	1	2	3	4	5		4
5. Command & Self Assertion	1	2	3	4	5		5
6. Support of Subordinates	1	2	3	4	5		6
7. Briefing Others	1	2	3	4	5		7
8. Knowledge of Trade/ Knowledge of Job When Employed out of Trade	1	2	3	4	5		8
9. Ability to Apply His/Her Knowledge	1	2	3	4	5		9
10. Adaptability	1	2	3	4	5		10
11. Initiative	1	2	3	4	5		11
12. Appearance and Bearing	1	2	3	4	5		12
13. Supervision	1	2	3	4	5		13
14. Ensuring Understanding of Assignments	1	2	3	4	5		14
15. Responsibility	1	2	3	4	5		15
16. Conduct	1	2	3	4	5		16
17. Learning From Experience	1	2	3	4	5		17

A Circle one of the following:
(1) I am a student
(2) I am a staff member

B Location of Junior Leaders' School - Circle One

1. CFB Borden
2. CFB Penhold
3. CFB Esquimalt
4. CFB Europe
5. CFB Halifax
6. CFB Summerside

C What suggestions do you have that would improve: (a) the current system of performance evaluation reporting, (b) Junior Leadership Assessment procedures?

D Feel free to comment on any aspect of this study by writing your comments on the back of this page.

THANK YOU FOR YOUR ASSISTANCE.

1378

TABLE B1

MEAN RATINGS OF DEGREE OF OBSERVABILITY BY STAFF AND STUDENTS OF THE 17 PERFORMANCE REQUIREMENTS (PR) OF LEADERSHIP

	STAFF (n=75)		STUDENTS (n=480)	
Performance Requirements (PR)	X	S.D.	X	S.D.
Preparation and Planning	4.25	.75	3.80	.90
Delegation	3.89	.79	3.78	.86
Performance Under Stress/ Pressure	3.56	.94	3.68	.91
Cooperation	3.90	.60	3.75	.66
Command and Self Assertion	3.88	.75	3.71	.81
Support of Subordinates	3.21	1.05	3.22	1.11
Briefing others	4.03	.81	3.74	.88
Knowledge of Trade/Knowledge of Job When Employed Out of Trade	3.51	1.00	3.22	1.07
Ability to Apply His/Her Knowledge	3.73	.90	3.62	.84
Adaptability	3.73	.87	3.64	.85
Initiative	3.88	.83	3.69	.83
Appearance and Bearing	4.43	.77	4.22	.93
Supervision	3.95	.92	3.70	.85
Ensuring Understanding of Assignment	3.67	.91	3.60	.86
Responsibility	3.96	.74	3.98	.85
Conduct	3.92	.89	4.09	.84
Learning from Experience	3.95	.86	4.00	.91

TABLE B2

RANK ORDER BY STAFF AND STUDENTS OF THE DEGREE OF OBSERVABILITY OF THE 17 PERFORMANCE REQUIREMENTS (PR) OF LEADERSHIP

Performance Requirements (PR)	Rank Order a) b) STAFF (n=75)	STUDENTS (n=480)
1. Preparation and Planning	2	5
2. Delegation	9	6
3. Performance Under Stress/ Pressure	15	12
4. Cooperation	8	7
5. Command and Self Assertion	10	9
6. Support of Subordinates	17	17
7. Briefing others	3	8
8. Knowledge of Trade/Knowledge of Job When Employed Out of Trade	16	16
9. Ability to Apply His/Her Knowledge	13	14
10. Adaptability	12	13
11. Initiative	11	11
12. Appearance and Bearing	1	1
13. Supervision	6	10
14. Ensuring Understanding of Assignment	14	15
15. Responsibility	4	4
16. Conduct	7	2
17. Learning from Experience	5	3

a) Spearman's rank order correlation = .87 $p<.0001$

b) Kendall's tau= .72 $p<.0001$

TABLE B3

AN RATINGS OF RELATIVE IMPORTANCE OF THE 17 PERFORMANCE REQUIREMENTS (PR) LEADERSHIP EFFECTIVENESS

	STAFF (n=78)		STUDENTS (n=506)	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.
eparation and Planning	2.72	3.01	4.28	3.83
legation	7.85	4.41	8.36	4.44
rformance Under Stress/ essure	10.21	4.87	7.00	4.92
operation	6.92	2.48	6.55	2.75
mmand and Self Assertion	6.87	4.45	6.64	4.52
pport of Subordinates	11.76	3.75	10.53	4.57
iefing others	6.59	3.63	8.84	3.95
owledge of Trade/Knowledge Job When Employed Out of ade	9.41	6.22	10.81	6.27
ility to Apply His/Her owledge	7.62	4.69	9.09	5.04
aptability	9.67	4.86	8.66	4.71
itiative	7.27	3.38	7.84	4.16
pearance and Bearing	13.26	3.38	12.86	4.81
pervision	6.88	3.11	8.84	4.24
suring Understanding of signment	10.12	3.72	9.96	4.05
sponsibility	8.41	4.58	7.73	4.36
nduct	12.82	3.37	11.76	4.48
arning from Experience	12.40	3.69	9.87	4.78

TABLE B4

STAFF AND STUDENTS RANK ORDER OF THE RELATIVE IMPORTANCE OF THE 17 PERFORMANCE REQUIREMENTS TO LEADERSHIP EFFECTIVENESS.

Performance Requirements (PR)	Rank Order[a,b] STAFF (n=78)	STUDENTS (n=506)
1. Preparation and Planning	1	1
2. Delegation	8	7
3. Performance Under Stress/ Pressure	13	4
4. Cooperation	5	2
5. Command and Self Assertion	3	3
6. Support of Subordinates	14	14
7. Briefing others	2	10
8. Knowledge of Trade/Knowledge of Job When Employed Out of Trade	10	15
9. Ability to Apply His/Her Knowledge	7	11
10. Adaptability	11	8
11. Initiative	6	6
12. Appearance and Bearing	17	17
13. Supervision	4	9
14. Ensuring Understanding of Assignment	12	13
15. Responsibility	9	5
16. Conduct	16	16
17. Learning from Experience	15	12

[a] Spearman's rank order correlation = .69 $p < .01$

[b] Kendall's tau = 0.52 $p < .002$

TABLE B5

RELEVANCE OF PRs TO LEADERSHIP ASSESSMENT FOR STAFF AND STUDENTS
(n=373)

	Main Reason for Recommended Dropping[1].				Frequency of Total Ratings	
	1	2	3	4		%
·eparation and Planning				X	12	1.51
·legation				X	32	4.04
·rformance Under Stress/ ·essure			X		32	4.04
›operation	X				61	7.69
·mmand and Self Assertion				X	25	3.15
ıpport of Subordinates		X			94	11.85
·iefing others			X		31	3.91
ıowledge of Trade/Knowledge Job When Employed Out of ·ade	X				197	24.84
›ility to Apply His/Her ıowledge				X	49	6.18
laptability				X	35	4.41
ıitiative			X		24	3.03
›pearance and Bearing			X		44	5.55
ıpervision				X	23	2.90
ısuring Understanding of ssignment			X		58	7.31
›sponsibility				X	24	3.03
›nduct			X		30	3.78
›arning from Experience	X				22	2.77
					n=793	100%

›gend

Not appropriate (valid) for JL Assessment.

Not observable within the training program.

Not practical in its present form.

Duplication of other ›Rs.

TABLE B6

FREQUENCY COUNT OF EXTENT OF OBSERVATION OF PIs STAFF (N = 75) AND STUDENTS (N = 480)

	Staff					Students				
	Never		Adequate		Frequently	Never		Adequate		Frequently
	1	2	3	4	5	1	2	3	4	5
1. Preparation and Planning	0	0	14	28	33	0	37	130	183	115
2. Delegation	1	0	22	35	17	0	29	156	196	109
3. Performance Under Stress/Pressure	1	8	27	26	13	1	46	156	179	98
4. Cooperation	5	13	33	20	4	37	159	133	134	17
5. Command & Self Assertion	0	2	20	38	15	0	25	172	239	64
6. Support of Subordinates	4	13	25	23	8	28	106	144	138	64
7. Briefing Others	0	1	19	28	27	0	30	175	164	111
8. Knowledge of Trade/Knowledge of Job When Employed out of Trade	1	10	29	20	15	29	91	161	145	54
9. Ability to Apply His/Her Knowledge	0	7	22	30	16	7	37	161	200	75
10. Adaptability	1	3	26	30	15	3	31	177	192	77
11. Initiative	0	4	19	34	18	0	34	162	202	82
12. Appearance and Bearing	0	0	13	17	45	5	15	90	121	241
13. Supervision	2	2	16	33	22	2	34	155	206	83
14. Ensuring Understanding of Assignments	0	2	21	31	14	0	46	176	183	75
15. Responsibility	0	0	22	34	19	0	25	102	210	145
16. Conduct	1	2	21	29	22	0	15	103	185	177
17. Learning from Experience	1	3	15	36	20	3	30	91	196	160

Appendix C

DEFINITIONS OF RECOMMENDED CRITICAL REQUIREMENTS

1386

DEFINITIONS OF CRITICAL REQUIREMENTS

DOES HE OR SHE

1. ASSUME RESPONSIBILITY

   a. Fulfills his/her duties consistently.

   b. Takes upon him/herself to do the tasks without any noticeable reluctance.

   c. Takes actions and accepts the responsibility for their consequences.

2. TAKE COMMAND AND ASSERT HIMSELF/HERSELF

   a. Asserts him/herself as leader.

   b. Demonstrates confidence in his/her ability to lead.

   c. Generates the respect of associates.

   d. Inspires the confidence of others.

3. PREPARE AND PLAN

   a. Considers resources and conditions.

   b. Prepares a realistic plan.

   c. Plans flexibly.

   d. Plans in detail to include the Who, What, When, Where and How of planning using logical analysis.

4. BRIEF OTHERS

   a. Provides others with the needed information to do the job.

   b. Gives clear, complete and concise orders demonstrating where appropriate.

   c. Organizes his/her orders logically.

   d. Ensures others have understood his/her instructions.

5. SUPERVISE

a. Checks to ensure task was being carried out as directed, and corrected where necessary.

b. Gives assistance and advice to subordinates throughout the task.

c. Delegates authority to subordinates for supervision of portions of the task as needed.

d. Retains overall control of the operation.

e. Issues new instructions to compensate for changes in the situation.

6. DISPLAY ADAPTABILITY/INITIATIVE

a. Steps in and takes action when the need arises.

b. Adapts by displaying imagination and flexibility of approach.

c. Develops new solutions when familiar methods fail.

d. Takes action on his/her own volition.

e. Makes timely decisions.

f. Demonstrates flexibility in thought and behaviour.

g. Tries new ideas to achieve results.

7. COMMUNICATE EFFECTIVELY

a. Employs confident, enthusiastic delivery.

b. Employs volume, tone and modulation appropriate to the situation.

c. Articulates clearly.

8. PERFORM EFFECTIVELY UNDER STRESS

a. Retains self-control under difficult conditions.

b. Maintains the pursuit of goals under difficult conditions.

c. Achieves acceptable results under physical/emotional stress.

9. CORRECTLY APPLY KNOWLEDGE

a. Demonstrates adequate knowledge of the tasks at hand.

b. Demonstrates his/her ability to utilize new knowledge and experience.

c. Profits by instruction.

d. Demonstrates by his/her actions a mastery of knowledge acquired.

e. Applies theory to specific situations or problems.

f. Reorganizes his/her approach or plan when needed.

10. COOPERATE WITH OTHERS

a. Works as a member of a team (not as an individual).

b. Cooperates with the associate who is in charge.

c. Makes the effort to amend strained personal relations in the group.

d. Deals tactfully with others.

e. Works with and through others.

f. Assists less capable associates in order to attain team

g. Willingly helps an associate in his work.

11. LEARN FROM EXPERIENCE

a. Converts mistakes into a source of learning.

b. Show receptiveness to advice and profits by it.

c. Admits his/her mistakes and accepts their consequences.

Appendix D

RECOMMENDED LEADERSHIP PERFORMANCE RATING SCALE

LEADERSHIP PERFORMANCE RATING SCALE

DID HE OR SHE:

1. ASSUME RESPONSIBILITY

a. Failed to assume assigned responsibilities.	2
b. Avoided assigned responsibilities	4
c. Accepted assigned responsibilities.	6
d. Readily accepted assigned responsibilities.	8
e. Readily accepted assigned responsibilities and sought additional responsibilities.	10

2. TAKE COMMAND AND ASSERT HIM/HERSELF

a. Failed to assert his/her command and as a result lost complete control of his/her subordinates.	2
b. Inadequately asserted his/her command and as a result lost some control of his/her subordinates.	4
c. Adequately asserted his/her command and as a result retained acceptable control of his/her subordinates.	6
d. Effectively asserted his/her command and retained good control of his/her subordinates.	8
e. Was exceptional in asserting his/her command and retained excellent control of his/her subordinates.	10

3. PREPARE AND PLAN

a. Because he/she failed to prepare him/herself he/she had no plans for the completion of tasks.	2
b. Prepared inadequate plans as a result of poor application of planning techniques.	4
c. Prepared adequate plans as a result of a satisfactory application of planning techniques.	6
d. Prepared careful plans as a result of a sound application of planning techniques.	8
e. Prepared excellent plans as a result of a thorough application of planning techniques.	10

1. HELP OTHERS

a. Failed to provide and check the information and instructions given to subordinates to ensure completion of tasks. 2

b. Provided information and instruction which permitted only partial completion of tasks and made inadequate checks of subordinates' understanding of such instructions. 4

c. Provided sufficient information and ensured subordinates' understanding of major points of instruction which permitted the satisfactory completion of tasks. 6

d. Provided information and ensured subordinates' general understanding of instruction which permitted a high level of task completion. 8

e. Provided information and tested the understanding of instruction given, ensuring subordinates knew in detail what they were to do, which permitted a high level of task completion. 10

3. SUPERVISE

a. Failed to supervise the activities of subordinates in the performance of assigned tasks. 2

b. Did not adequately supervise the activities of subordinates in the performance of assigned tasks. 4

c. Adequately supervised the activities of subordinates in the performance of assigned tasks. 6

d. Effectively supervised the activities of subordinates in the performance of assigned tasks. 8

e. Was exceptional in his/her supervision of subordinates in the performance of assigned tasks. 10

4. DISPLAY ADAPTABILITY/INITIATIVE

a. Failed to make decisions or take action on his/her own volition as a result of his/her failure to adapt to new problems or situations. 2

b. Delayed required decisions and actions as a result of his/her inadequate ability to adapt to new problems or situations. 4

c. Made necessary decisions and took appropriate action as a result of acceptable ability to adapt to new problems or situations. 6

d. Made prompt decisions and took early action as a result of his/her effective ability to adapt to new problems or situations 3

e. Made prompt decisions and took immediate action as a result of his/her superior adaptation to new problems or situations. 10

7. COMMUNICATE EFFECTIVELY

a. Failed to express him/herself confidently by using poor oral communication skills. 1

b. Inadequately expressed him/herself by using few oral communication skills. 2

c. Adequately expressed him/her self by using several oral communication skills. 3

d. Confidently expressed him/herself by using most oral communication skills 4

e. Displayed excellent oral communication skills. 5

8. PERFORM EFFECTIVELY UNDER STRESS

a. Became entirely ineffective under periods of stress. 1

b. Allowed periods of stress to interfere with his/her effectiveness. 2

c. Maintained an acceptable level of effectiveness in spite of periods of stress. 3

d. Worked effectively under periods of stress. 4

e. Was very effective under periods of stress. 5

9. CORRECTLY APPLY KNOWLEDGE

a. Could not apply his/her knowledge with assistance. 1

b. Required frequent guidance in applying his/her knowledge. 2

c. Required little assistance in applying his/her knowledge. 3

d. Applied his/her knowledge without assistance. 4

e. Applied his/her knowledge with a high degree of skill. 5

10. COOPERATE WITH OTHERS

a. Made no effort to cooperate with others, and by his/her failure to work as a team member lowered the effectiveness of the group. 1

b. Made little effort to cooperate with others, and by only occasionally working as a team member, lowered the effectiveness of the group. 2

c. Made an adequate effort to cooperate with others, and by acceptably working as a team member, added to the effectiveness of the group. 3

d. Made a positive effort to cooperate with others, by working effectively as a team member, thereby increasing the overall effectiveness of the group. 4

e. Made an exceptional effort to cooperate with others and by working very effectively as a team member, greatly enhanced the overall effectiveness of the group. 5

11. LEARN FROM EXPERIENCE

a. Refused to admit mistakes. 1

b. Admitted mistakes with reluctance. 2

c. Acknowledged mistakes, accepted advice and utilized the opportunity to learn from the experience. 3

d. Readily admitted mistakes, sought advice, and utilized the opportunity to learn from the experience. 4

e. Volunteered information about his/her mistakes, sought advice and utilized the opportunity to learn from the experience. 5

FROM TASK INVENTORIES TO PERFORMANCE APPRAISAL FORMS

Roy R. Van Cleve

In a recent study of the nation's 500 largest industrial firms,[1] of those which stated they paid their salaried employees based upon performance (90% of those responding) only 17% used appraisal devices which were actually job performance oriented. For the others, other rating factors, primarily personal characteristics, predominated.

In another recent study of a "promotional potential" (performance appraisal) form, the authors stated:

> The ideal PP would probably be one that is class specific. Through a thorough job analysis, specific KSA's could be identified. The KSA's would then be examined, and those KSA's that could be demonstrated by on-the-job observation would be included in the PP. Although such a PP would be valid, the time involved in preparation might be prohibitive.[2]

While a third study offers the following:

> Organizations today do not have a choice of whether or not to examine their present performance appraisal systems; the choice actually is between the alternatives available to them.[3]

Indeed, the Tenth Circuit Court of Appeals, in the case of Brito vs. Zia, held that the Zia Company's employee performance evaluations, used for lay-off, were in fact tests under the meaning of EEOC Guidelines, and were thus subject to empirical validation as required by EEOC.

Additionally, as Holley, et al, state,

> ...any new performance appraisal system must be tailored to the specific needs and objectives of a particular organization's idiosyncrasies and organizational climate. Moreover, if the ratings are to be used as a basis for personnel decision making, EEOC regulations require a thorough job analysis, specification of job-related rating factors, standardized administration and non-discriminatory results.[4]

The authors of these articles do not, however, leave us without some recommendations for more meaningful performance appraisal systems. The relevant paths to improvement offered by Winstanley are:

Train managers, especially in feedback.

Develop job-specific and success-relevant scales. (This will undoubtedly mean more and different appraisal methods in a given company.)

To aid acceptance, allow employee participation, at least in initial design.

[Underlining provided for emphasis]

While Beckett and Dodsworth offer the following for developing an effective appraisal scheme:

> --it should be designed to meet specific aims;
> --line management should be involved in, and responsible for, the design and implementation of the scheme using specialist personnel assistance as necessary;
> --assessment should be against objective targets agreed with the job holder and quantified where possible;
> --managers should be trained in the purpose and procedures of the scheme as well as in the necessary interviewing skills.[6]

An answer to the problems set forth, and in keeping with the recommendations, a solution for the development of a task-based performance appraisal system and other manpower related areas has been developed in recent years.

Figure 1 clearly shows that the task-based, time ordered, task inventory, survey systems approach to occupational analysis, when combined with the Comprehensive Occupational Data Analysis Programs (CODAP), has broadened the impact of "job analysis" to the point where it now touches every facet of human resources management and provides the thorough analysis of the processed data necessary for the task-based job descriptions to be utilized in all of these functions:

1. Organizational and manpower planning
2. Position classification
3. Career planning
4. Personnel selection, assignments, and utilization
5. Job restructuring, evaluation, and pricing
6. Training
7. Performance Evaluation
8. Qualification standards
9. Equipment utilization
10. Research into mental and physical demands, job satisfaction and personnel retention.

COMPARATIVE ATTRIBUTES OF OCCUPATIONAL ANALYSIS SYSTEMS

	Manual Systems			Computer Assisted Systems	
Attribute of[b] System	Engineering Job Analysis	Functional Job Analysis	Check List Job Analysis	JAV - DECAL[a] FJA	CODAP Task Based, Time Ordered
Measurement specificity & accuracy	VG	F - G	F - VG	P	VG in Armed Forces; Needs verification in private sector.
Utility for generalizing	P	P - F	F - VG	P	VG
Relative Cost	Very costly	Costly	Economical	Economical	Economical
Relative speed & capacity	Very slow & very limited	Slow & limited	Slow & high capacity	Fast & high capacity	Fast & high capacity
Exhaustive-ness	P	F	P - VG	P	G
Application utility					
Certification standards development	G (limited)	P	P	P	EX
Job engineering	EX	P	P - VG	P	EX
Man-job matching	P	G	F - VG	P - F	EX
Career ladder design	P	F	P - F	P	EX
Performance evaluation	P - G	P - F	F - G	P	EX
Curriculum development	G - VG	P	P - F	P - F	EX
Aggregate Assessment	Limited utility	Limited utility	Limited utility	Limited utility	High potential Needs testing

Figure 1

[a] Both JAV and DECAL are experimental.

[b] Appendix 1 provides a discussion of the attributes of the systems.

P = Poor; F - Fair; G = Good; VG = Very Good; EX = Excellent.

Source: Occupational Analysis for Human Resource Development: A Review of Utility of the Task Inventory, Research Report Ms., 25 OCMM, Dept. of Navy, Brian E. Moore, PhD; University of Texas; 1976.

The systems flow chart, Figure 2, depicts the TI/CODAP process we used in our project. Each of the four job families in the project included the jobs, currently described in the state classification plan, which an entry-level job incumbent could reasonably expect to encounter in a normal career progression. The population of current job incumbents, numerically, by job type and pay range, was examined to ensure that all of the then current positions were included in at least one job family.

In preparation for inventory construction, each job family was researched in depth. A detailed review of the literature was conducted, which included a study of the current state classification plan and the approved job descriptions; an examination of the task inventories obtained from the Task Inventory Exchange (TIE) and from other sources which had not contributed their inventories to TIE; and an extensive study of research reports and publications.

Upon completion of review of the literature, the analysts conducted observation and interview sessions with selected job incumbents, supervisors, and training-oriented subject matter experts.

Following that, the initial technical committee meetings were conducted. In those meetings, the analysts presented their initial concepts of the job families and received feedback from the committees, ensuring them that the research and observation and interview sessions had uncovered all significant job factors and that the analysts were on the right track and ready to proceed with inventory construction.

Upon completion of the inventory construction and after the initial printing of the inventories, a second set of technical committee meetings was held. In these meetings, each item to be included in the inventory was reviewed by the committee members. Any item which was considered to be questionable, inappropriate, or invalid was researched in greater depth by the job analysts.

On all items which were considered questionable, inappropriate, or invalid, the analysts once again contacted either the job incumbent, supervisor, or subject matter expert, or all of them, and when necessary rewrote the item or task statement in a valid format.

Once all of the questionable items had been resolved, the inventories were printed in a format which facilitated incumbent responses during the inventory administration.

The inventories were then administered to proctored groups of employees in order to ensure individual, vis-a-vis group, responses to the inventory.

To verify the accuracy of responses on the standard answer sheets used, each booklet was compared to its attendant set of answer sheets to ensure completeness of information in the background, equipment list, and job survey data sections, and to verify that when a level of involvement was

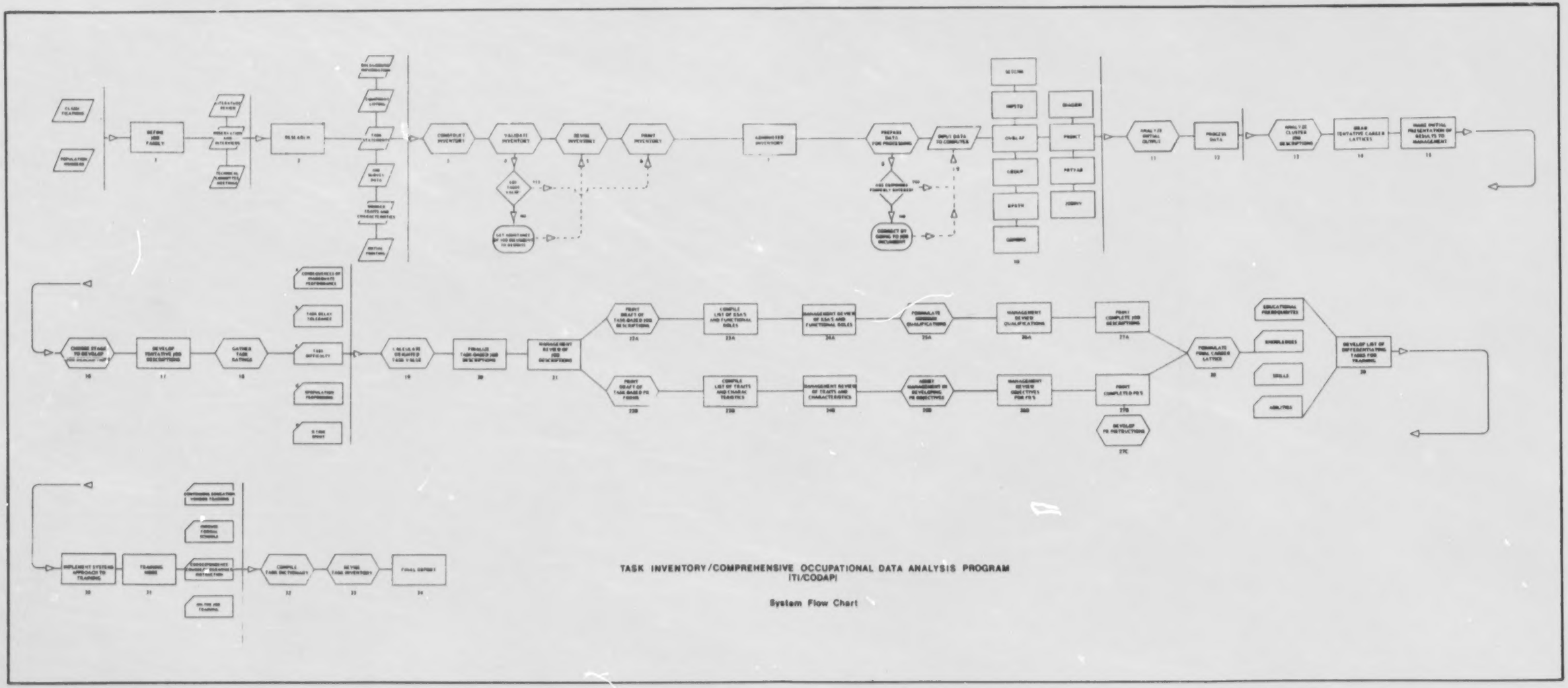

TASK INVENTORY/COMPREHENSIVE OCCUPATIONAL DATA ANALYSIS PROGRAM
(TI/CODAP)
System Flow Chart

Figure 2

Using the selected stages' job descriptions, the analysts and technical committee members developed tentative job descriptions which indicated the appropriate level of involvement in each task for the learner, worker, lead work, first line, and staff supervisor. They then determined which of the current classifications should be included in each of the new job descriptions. Finally, all of the tasks which were inappropriate to that particular job description for those classifications and pay grades were eliminated from the list.

These tentative job descriptions provided information about the percentage of population which performed those tasks and the relative amount of time spent in the performance of each task. This was viewed as the job incumbent's input to the relative importance of each task in the performance of their jobs. Management was then asked to rate the task's relative importance on these factors -- consequences of inadequate performance, task delay tolerance, and task difficulty using the instructions shown in Figure 3.

The immediate CONSEQUENCES OF INADEQUATE PERFORMANCE (CIP) are much more serious for some tasks than for others. For example, the failure to "Furnish data necessary to take seizure action on delinquent taxpayers" could immediately allow the taxpayer to dispose of any materials we could seize or allow the taxpayer to leave undetected, which would ultimately result in the loss of state income. This failure would be considerably more serious than the failure to "Review completed documents for accuracy of spelling, punctuation, and grammar."

TASK DELAY TOLERANCE (TDT) is defined as how much delay can be tolerated before the task must be performed. It can be described as the time difference between when the need for a task to be performed becomes evident and when the task performance must begin.

Tasks with less Delay Tolerance must be performed without delay and would be given a high number rating. Tasks with more Delay Tolerance (that is, time could be spent without jeopardy researching House bills or Attorney General opinions before the task has to be performed) would be given a lower number rating.

The criterion for determining TASK DIFFICULTY (TD) is the average amount of time required to learn the techniques utilized in the department.

Upon completion of the ratings by management personnel, weighted task values were calculated. A policy decision had been made that each of the management-rated factors was of equal importance, therefore, the raw weighted task value (WTV) was obtained by multiplying CIP x TDT x TS = WTV. To analyze the impact of incumbent input, percent of population performing (PP) and percent of time spent (PT) on WTV, PP x WTV and PT x WTV were calculated and then PP x PT x WTV = TWTV was determined (Figure 4).

When the calculations for weighted task value had been completed, the analysts and technical committee members met to finalize the task-based job descriptions. During these meetings, the expected level of involvement

CONSEQUENCES OF INADEQUATE PERFORMANCE (CIP)

RATING SCALE	BENCHMARK RATINGS
1. Mildly Embarrassing	1. Make Xerox copies
2. Slight Impact	2. Fold/Stuff correspondence for mailing
3. Greater Impact, but not Serious	3. Maintain alpha file
4. Not very Serious	4. Prepare Employee Performance Reviews
5. Fairly Serious	5. Answer routine inquiries in person/ on the phone
6. Serious Impact	6. Prepare deposit vouchers for processing twice daily
7. Very Serious Impact	7. Contact taxpayers in person to secure payment for "hot checks"
8. Almost Disastrous	8. Calculate amounts due the State and City from delinquent taxpayers for Attorney General
9. Disastrous	9. Furnish certification of taxpayer liability to Attorney General for bankruptcy

TASK DELAY TOLERANCE (TDT)

RATING SCALE	BENCHMARK RATINGS
1. Within the Biennium	1. (No benchmark was established)
2. Within the Fiscal Year	2. Prepare grant proposals
3. Within the Quarter	3. Prepare employee performance reviews
4. Within the Month	4. Furnish certification of taxpayer liability to Attorney General for bankruptcy
5. Within Two Weeks	5. Examine claims submitted by state agencies for proper allocation
6. Within the Week	6. Review weekly activity reports to check progress of scheduled work
7. Within the Day	7. Log daily readings on postal machines
8. Within the morning or afternoon	8. Prepare deposit vouchers for processing daily
9. Instantaneous - Must Do Immediately	9. Answer telephone inquiries

TASK DIFFICULTY SCALE (TD)

1. Short demonstration only -- less than one hour
2. Over one hour up to and including eight hours (one day)
3. Over one day up to and including one week
4. Over one week up to and including one month
5. Over one month up to and including three months
6. Over three months up to and including six months
7. Over six months up to and including one year
8. Over one year up to and including two years
9. Over two years

Figure 3

08/04/76 JOB DESCRIPTION WEIGHTED TASK VALUES PAGE 1

ACCOUNTS EXAMINER I

TASK NUMBER	TASK DEFINITION	AVG CIP	AVG TDT	AVG TDF	RATE WEIGHT	PCT PERF	PERF WEIGHT	PCT TIME	TIME WEIGHT	TOTAL WEIGHT
DUTY:	TAXPAYER SERVICE									
011	ANSWER ROUTINE INQUIRIES IN PERSON/ON THE PHONE	4.33	6.55	4.55	129.03	90	116.12	2.14	276.12	248.49
014	ANSWER ROUTINE INQUIRIES CONCERNING TAX PROBLEMS BY TELEPHONE	3.88	6.77	2.88	75.62	93	70.32	2.07	156.53	145.56
166	FURNISH INFORMATION OF A COMPLEX NATURE TO TAXPAYERS CONCERNING LAWS, REGULATIONS, AND PROCEDURES BY TELEPHONE	3.00	3.66	4.66	51.16	59	30.18	1.81	92.59	54.62
013	ANSWER ROUTINE INQUIRIES CONCERNING TAX PROBLEMS IN PERSON	3.00	4.77	3.00	42.93	47	20.17	1.31	56.23	26.42
022	ASSIST TAXPAYERS IN PREPARING TAX REPORTS AND RETURNS	3.00	5.00	2.55	38.25	63	24.09	1.36	52.02	32.76
012	ANSWER ROUTINE INQUIRIES CONCERNING TAX PROBLEMS BY CORRESPONDENCE	2.55	3.77	2.66	25.56	91	23.25	2.20	56.23	51.15
285	PREPARE TAXPAYER CHANGE REQUEST (TCR) FOR CHANGES IN TAXPAYER'S ACCOUNT	2.77	3.55	2.00	19.66	72	14.15	1.72	33.81	24.33
165	FURNISH INFORMATION OF A COMPLEX NATURE TO TAXPAYERS CONCERNING LAWS, REGULATIONS, AND PROCEDURES BY CORRESPONDENCE	2.33	2.88	2.88	19.32	57	11.01	1.91	36.90	21.02
167	FURNISH INFORMATION OF A COMPLEX NATURE TO TAXPAYERS CONCERNING LAWS, REGULATIONS AND PROCEDURES IN PERSON	2.44	2.44	2.88	17.12	27	4.62	1.40	23.96	6.46
155	EXAMINE TAXPAYERS' RECORDS TO DETERMINE IF BOND OR SECURITY REFUND REQUEST IS JUSTIFIED.	2.44	2.77	2.33	15.73	20	3.14	1.64	25.79	5.14
016	MAKE ARRANGEMENTS FOR HEARINGS AS REQUESTED BY TAXPAYER	1.00	1.55	.44	.68	23	.15	1.65	1.12	.24
189	ISSUE TEMPORARY TRIP PERMITS TO FUEL TRUCKS	.55	1.00	.55	.30	10	.03	2.49	.74	.07
021	ASSIST TAXPAYER IN PREPARING APPLICATIONS FOR PERMITS, LICENSES AND OTHER RELATED DOCUMENTS	1.44	2.66	.00	.00	25	.00	1.11	.00	.00
	SPEAK BEFORE VARIOUS PUBLIC GROUPS ABOUT MATTERS RELATING TO THE OPERATIONS OF THE DEPARTMENT	.22	.00	.00	.00	1	.00	.26	.00	.00

1972

Figure 4

1403

AX COLLECTION

1404

in each task for each level of position was defined, duty categories and titles were agreed upon, and the tasks were rank ordered in each duty category. Task statements were selected as performance review items for inclusion in the performance review forms which were to be developed, and once again, those tasks which were inappropriate at the selected levels were eliminated.

Work in the previous step resulted in the basic job descriptions. Following management concurrence with the basic job descriptions, work progressed in parallel on the development of the total job description and the performance review forms.

In order to facilitate the work of the analysts and supervisory personnel involved in the development of the total products, sufficient copies of drafts were printed for distribution as working papers, and comments were solicited directly on the forms.

Using the Department of Labor definitions of knowledges, skills, and abilities, a comprehensive list of KSA's was compiled for each job category, each level within those categories, and each task within each level of the categories. Functional role statements were developed to show how the tasks related to the functions of the total job. These statements were expressed in terms of data, people, things and concepts, and the amount of supervision given and received.

In addition to the performance of the tasks which comprise the job, possession of certain basic or unlearned qualities is also required of the worker. These basic abilities are known as worker traits or worker characteristics. These characteristics are used to aid in the selection of new workers and in the reassignment, reclassification, and promotion of current workers. They identify qualities which are not clearly measurable in the performance of a single task.

The traits and characteristics chosen for inclusion in our task-based performance review form were dictated by two major considerations.

First, the trait or characteristic had to have a personal manifestation of job-related mannerisms, aptitudes, or attitudes in groups of tasks or the total job, instead of a knowledge, skill, or ability which was manifest in a single task.

Second, the trait or characteristic had to be identifiable and measurable for evaluation.

Following the completion of KSA's and Functional Roles for the job descriptions and worker traits and of characteristics for the performance review forms, the results were presented to management for their review, input, and concurrence.

Upon receipt of management's concurrence of the list of KSA's and functional roles, the job analysts developed standard lists of basic requirements for entry-level hiring and for promotional opportunities, which were task-based and directly linked to the previously defined critical knowledges, skills, and abilities.

The minimum qualifications were then submitted to management for their review, input, and concurrence.

Approval of the minimum qualifications completed the elements of the job descriptions. The functional roles (job summary), the rank-ordered within duties task statements, the compilation of KSA's and minimum qualifications were then ready to be printed in a single management-usable form. Figure 5 is the complete job description for the Auditor I developed by this process.

After approval of the tasks and the traits and characteristics to be included in the performance review forms and having left an open section for the incorporation of MBO results at a later date, the forms were then printed. Figure 6 is an example of the complete Auditor I Performance Review Form.

Concurrent with the development of the performance review forms, an instructional manual and desk-top marking aid were developed. The manual was designed to serve as an in-depth reference guide and the basis for an instructional period to be conducted for all marking supervisors. The desk-top marking aid was designed for use as a ready reference during the marking of the forms.

When all of the job descriptions of a job family had been approved, the interrelationships of the jobs within the family were discernable. An analysis of the overlap of tasks between jobs showed which jobs provided readily identifiable career progression routes.

Once the career lattice was approved, lists of differentiating tasks were developed. These lists were used for two prime purposes. First, they formed the basic lists of tasks with their attendant knowledges, skills, and abilities which had to be mastered by the job incumbent prior to reclassification or promotion. Second, the task lists, with their attendant knowledges, skills, and abilities, and educational prerequisites provided the basis for formulating the technical training aspects of the organization's systems approach to training.

The training mode selected for each task is a policy decision which is based upon which of the training modes is the most cost-effective in obtaining the desired KSA's and educational prerequisites for the trainees.

The vast amounts of data collected for each of the task statements during the project were then compiled into a task dictionary, using the format shown below:

Figure 5

AUDITOR I

Summary of Job

Performs beginning professional auditing work as a trainee for the State Comptroller. Work involves assisting lead auditors in determining an audit program and conducting an audit. Work relationship to data, people and things is assisting in analyzing and compiling data, assisting in instructing and advising individuals and operating adding machines/calculators. Works under close supervision of a lead auditor with some latitude for independent judgment after training on routine assignments. Completed work is subject to review.

TASK STATEMENT	LEVEL	PRI
A. PRE-AUDIT		
1. ASCERTAIN STATUS OF ACCOUNT THROUGH TAXPAYER HISTORY	2	
2. REVIEW PERTINENT TAX LAWS, REGULATIONS AND LEGAL OPINIONS IN PREPARATION OF AUDIT	2	
3. INITIATE INQUIRIES OF TAXPAYERS' ACCOUNTS ON TERMINAL	2	X(combined 1 - 6)
4. DETERMINE TYPE OF BUSINESS IN PREPARATION OF AUDIT	1	
5. ANALYZE APPROPRIATE INDUSTRY TO PREPARE FOR AUDIT	1	
6. REVIEW PRIOR AUDIT WORK PAPERS	1	
B. DETERMINE AUDIT PROGRAM		
7. INSPECT PLANT FACILITIES AT AUDIT LOCATION	1	
8. DETERMINE/IMPLEMENT AUDIT SAMPLING TECHNIQUES	1	X
9. TEST REPORTING (ACCRUAL) PROCEDURE AND ANALYZE RESULTS	1	
C. CONDUCT AUDIT		
10. APPLY TAX STATUTES, COURT CASES AND ADMINISTRATIVE RULINGS DURING CONDUCT OF AUDIT	1	
11. PREPARE FIELD AUDIT SUMMARY OF DATA AND INFORMATION COLLECTED DURING AN AUDIT	1	
12. REVIEW GENERAL LEDGER FOR PERTINENT ACCOUNTS	1	
13. ANALYZE PERTINENT ACCOUNTS	1	
14. TEST SOURCE DOCUMENTS, SUBSIDIARY LEDGERS	1	
15. TRACE ENTRIES TO SOURCE DOCUMENTS	1	X
16. DOCUMENT AUDIT EXCEPTIONS (SCHEDULES, ETC.)	1	X (combined with 11 & 19)
17. COMPILE AUDIT RESULTS	1	
18. MAKE PROJECTIONS INTO PERIODS NOT EXAMINED IN DETAIL	1	
19. SUMMARIZE AUDIT RESULTS ON WORK PAPERS	1	X
20. CONDUCT TAXPAYER AUDIT CONFERENCE	1	
21. ANALYZE DATA TO DETERMINE IF AUDIT WOULD BE UNPRODUCTIVE IF PURSUED	2	
22. ANALYZE CIRCUMSTANCE OF BUSINESS TO DETERMINE IF IMMEDIATE COLLECTION IS NECESSARY (AFTER PROBATION PERIOD)	1	
23. COLLECT TAX DUE AND ANY PENALTY AND INTEREST INCURRED		
D. SELECT AUDIT		
24. IDENTIFY VALID AUDIT LEADS FOR SPIN-OFF AUDITS	2	X
E. TAXPAYER SERVICE CONTACT		
25. ANSWER ROUTINE INQUIRIES IN CONNECTION WITH AUDIT	1	
26. ASSIST TAXPAYERS IN PREPARING TAX REPORTS AND RETURNS	1	
27. ASSIST TAXPAYER IN PREPARING APPLICATIONS FOR PERMITS, LICENSES AND OTHER RELATED DOCUMENTS	1	
28. EXPLAIN PURPOSE AND PROCEDURES OF AUDIT TO TAXPAYER	1	
F. ADMINISTRATIVE		
29. DISCUSS PROCEDURES AND FINDINGS OF AUDIT IN PREPARATION FOR AND AT HEARINGS	1	
30. PREPARE TAXPAYER CHANGE REQUEST (TCR) FOR CHANGES IN TAXPAYER'S ACCOUNT	1	
31. IDENTIFY NEED FOR TRAINING	2	
32. COMPOSE ROUTINE WRITTEN COMMUNICATIONS	1	X
33. EVALUATE EFFECTIVENESS OF TRAINING PROGRAMS	1	
34. ATTEND TRAINING SCHOOLS, MEETINGS, CONFERENCES	2	
35. PREPARE TRAVEL VOUCHERS FOR REIMBURSEMENT OF EXPENSES INCURRED	2	
36. PREPARE TIME AND INVENTORY REPORTS	2	X
37. UPDATE DEPARTMENTAL INFORMATION, PROCEDURES AND POLICY MANUALS	2	
38. SUGGEST CHANGES IN EXISTING PROCEDURES FOR EFFICIENCY	2	

CRITICAL KNOWLEDGES, SKILLS AND ABILITIES

1. ABILITY TO ANALYZE INFORMATION
2. ABILITY TO COMPREHEND AND RETAIN TECHNICAL MATTER
3. ABILITY TO MAKE DECISIONS
4. KNOWLEDGE OF ACCOUNTING
5. SKILL IN USE OF ENGLISH LANGUAGE TO WRITE CLEARLY AND DISTINCTLY
6. SKILL IN ORGANIZING RESULTS ON AUDIT WORK PAPERS
7. ABILITY TO COMMUNICATE WITH OTHERS ORALLY WITH POISE, VOICE CONTROL AND CONFIDENCE
8. SKILL IN APPLYING TAX LAWS AND REGULATIONS TO SPECIFIC PROBLEMS OR SITUATIONS
9. ABILITY TO DEAL WITH STRESSFUL SITUATIONS
10. ABILITY TO FOLLOW DEPARTMENT PROCEDURES AND POLICIES
11. ABILITY TO LISTEN AND RETAIN INFORMATION

MINIMUM QUALIFICATIONS

ENTRY LEVEL

MINIMUM BBA OR 120 COLLEGE HOURS WITH 12 HOURS IN ACCOUNTING/AUDITING OR TWO YEARS OF RESPONSIBLE AUDITING/ACCOUNTING EXPERIENCE MAY SUBSTITUTE FOR 30 HOURS COLLEGE CREDIT, LIMITED TO FOUR YEARS OF EXPERIENCE FOR 60 HOURS OF COLLEGE CREDIT.

AUDITOR I

TASK-BASED PERFORMANCE REVIEW

I. Tasks

Expected Level of Involvement
(1) Assist (2) Do (3) Do & Supervise (4) Supervise

A. PRE-AUDIT

1. CONDUCT PRE-AUDIT RESEARCH AND REVIEW (2)	0 1 2 3 4 5 6 7 8 9

B. DETERMINE AUDIT PROGRAM

1. DETERMINE/IMPLEMENT AUDIT SAMPLING TECHNIQUES (1)	0 1 2 3 4 5 6 7 8 9

C. CONDUCT AUDIT

1. DOCUMENT AUDIT EXCEPTIONS (SCHEDULES, ETC.) (1)	0 1 2 3 4 5 6 7 8 9
2. CONDUCT TAXPAYER AUDIT CONFERENCE (1)	0 1 2 3 4 5 6 7 8 9
3. SUMMARIZE AND REPORT ON AUDIT FINDINGS (1)	0 1 2 3 4 5 6 7 8 9

D. SELECT AUDIT

1. IDENTIFY VALID AUDIT LEADS FOR SPIN-OFF AUDITS (2)	0 1 2 3 4 5 6 7 8 9

F. ADMINISTRATIVE

1. COMPOSE ROUTINE WRITTEN COMMUNICATIONS (1)	0 1 2 3 4 5 6 7 8 9
2. PREPARE TIME AND INVENTORY REPORTS (2)	0 1 2 3 4 5 6 7 8 9

II. Worker Traits and Characteristics

1. TACT	0 1 2 3 4 5 6 7 8 9
2. DECISIVENESS	0 1 2 3 4 5 6 7 8 9
3. DEPENDABILITY	0 1 2 3 4 5 6 7 8 9
4. INITIATIVE	0 1 2 3 4 5 6 7 8 9
5. REASONING ABILITY	0 1 2 3 4 5 6 7 8 9
6. FLEXIBILITY	0 1 2 3 4 5 6 7 8 9
7. EFFECTIVE COMMUNICATIONS	0 1 2 3 4 5 6 7 8 9
8. INTERPERSONAL RELATIONSHIPS	0 1 2 3 4 5 6 7 8 9
9. ABILITY TO DEAL WITH STRESSFUL SITUATIONS	0 1 2 3 4 5 6 7 8 9

III. Objectives

	1	2	3	4	5	6	7	8	9
Overall Evaluation of Employee	☐	☐	☐	☐	☐	☐	☐	☐	☐
Distribution of Marks for all Employees at this Level	☐	☐	☐	☐	☐	☐	☐	☐	☐

Recommendation	Yes	Group	Step	Effective Date
Merit Increase	☐	____	____	_____
Reclassification	☐	____	____	_____
Promotion	☐	____	____	_____

Narrative (Fully explain all adverse marks or recommendations)

Endorsements
I discussed this report with this employee on the date following my signature

Immediate Supervisor _____ Date _____

I have reviewed this report and offer the following comments:

2nd Echelon Supervisor _____ Date _____

The foregoing Performance Review has been discussed with me, and I wish to add the following comments:

Employee _____ Date _____

Figure 6

TASK DICTIONARY FORMAT

A. Task location by inventory
B. Task location by job specification
C. Level of involvement: Assist; Do; Do/Supervise; Supervise by %
D. % of personnel performing by job specification
E. % of time spent
F. Consequence of error
G. Delay tolerance
H. Task difficulty
I. Weighted task value
J. Conditions under which performed
K. Criteria for performance
L. Knowledges
M. Skills
N. Abilities
O. People
P. Data
Q. Things
R. Concepts
S. Formal school
T. Correspondence School
U. On-the-job training
V. Training course location(s)
W. Skills inventory item(s)
X. Performance Review Item(s)

The final step in the initial iteration of the project was the revision of the task inventories by the inclusion of all new and added tasks so that the updating of job descriptions in the future, utilizing the TI/CODAP approach, can be readily accomplished with a minimum of effort.

RESULTS

With completion of the initial iteration of work, the Comptroller of Public Accounts, State of Texas, has established an ongoing job analysis program, facilitated by the Comprehensive Occupational Data Analysis Program (CODAP). The results to date are:

1. Task-based job descriptions for 96 percent of the job incumbents.

2. Educational prerequisites, knowledges, skills and abilities qualifications established for each job description.

3. Career Lattices which clearly delineate upward mobility pathways for the three major job families.

4. Task-based performance review forms for each job description in the major job families.

5. The basic information for the implementation of a systems approach to training.

An additional employee benefit will be provided when employees are provided with copies of their current job descriptions, complete with KSA's, the performance review forms to be completed on them, and a copy of their job family's career lattice. In addition to that information, the employees will also be provided the job description(s) of all of the jobs in the career lattice to which they can routinely advance in a step-by-step progression. The performance review forms for those positions, along with a list of differentiating tasks with their attendant knowledges, skills and abilities, will also be provided. A part of the differentiating task

list will be instructions on where the differentiating task KSA's can be acquired, either by continuing education, consultant/vendor training, in-house formal schools, correspondence courses/programmed instructions, or on-the-job training. This information will allow the employees to plan their career progression and will serve as a tool for supervisory consultation.

As the work progressed in this project, the details of the Applied Job Analysis systems approach evolved.

Referring to Figure 7, it can be seen that as the "Sequential Functions" were executed, the analysts were able to "Determine" the "Individual Activity" and develop the elements of the basis of "Who, What, Where and How Much Time," the levels of involvement to identify "Learners, Workers, First Line Supervisors and Staff Supervisors," and how each category related to "Data, Things, People and/or Concepts." This "Task Inventory Data" enabled the analysts to spell out the basic "Job Requirements" and, as the work evolved, provided a "Revised Task Dictionary."

With the basis for the job requirements spelled out by the inventory data, the analysts were able to "Describe" the "Job Description(s)" in terms of the "Percentage of the Population Performing, Consequence of Inadequate Performance, Task Delay Tolerance and Task Difficulty" as well as percent of time spent, and arrive at "Weighted Task Value" job descriptions to be utilized in evaluating "Job Performance" with task-based "Performance Evaluation Review" forms.

When the job descriptions were complete, it was then possible to "Develop" a "Classification System" which contains graphically depicted career progression routes in the "Career Ladders/(Lattices)," have realistic "Qualification Standards" upon which to base "Recruitment Goals" and validate "Selection Devices" and "Performance Standards." This allowed for the establishment of viable "Staffing Criteria Goals" in which the total work forces' "Employment" supports "Staffing and Organizational Reports" which demonstrate an efficient and effective organization in action.

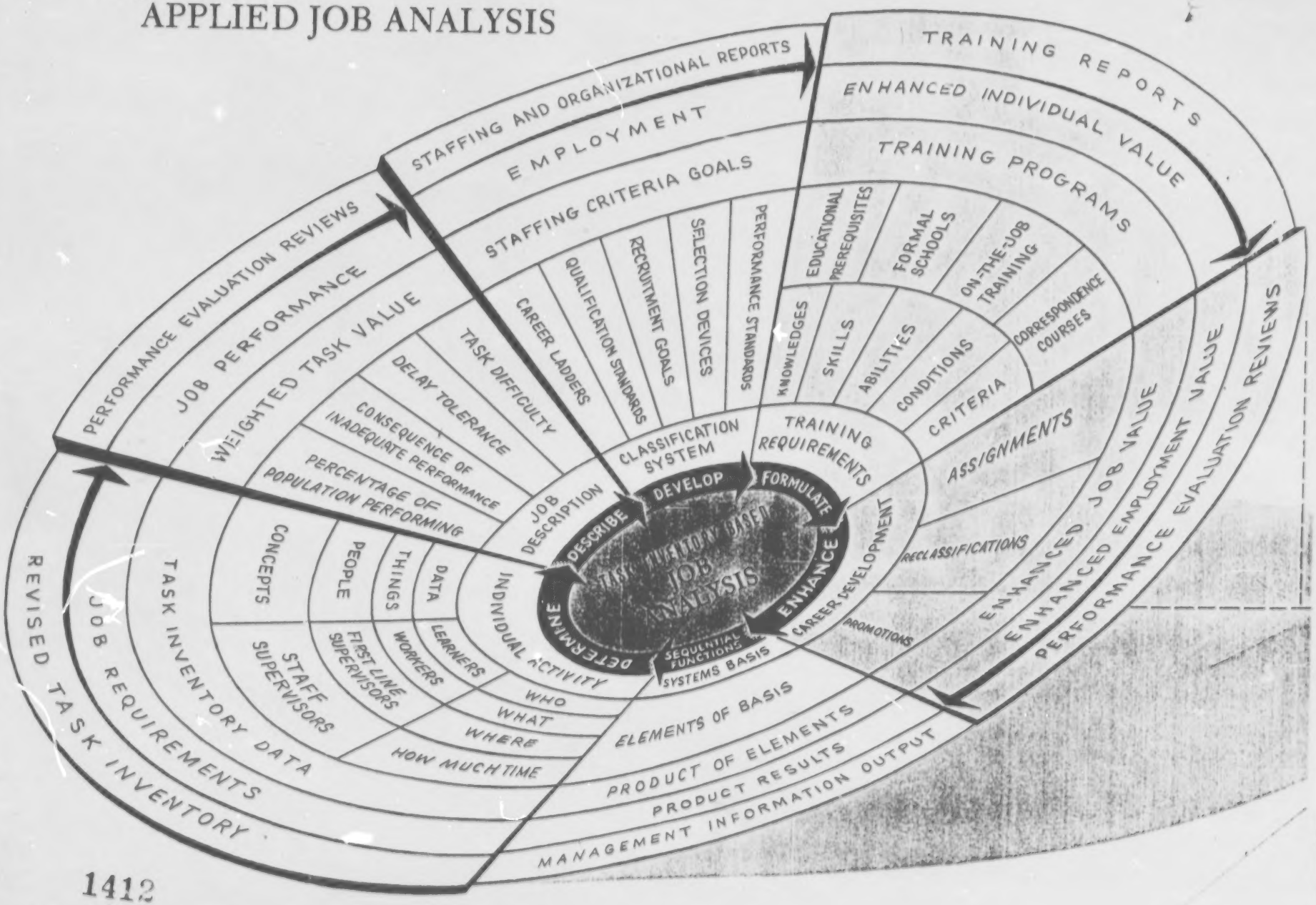
With the job descriptions and classification system as a solid basis, training specialists are able to "Formulate" the "Training Requirements" based upon precisely-stated requirements for "Educational Prerequisites, Knowledges, Skills and Abilities" in order for the employees to be able to perform under given "Conditions" to meet established training "Criteria." Taking all of these elements into consideration, decisions as to the most cost-effective mode of training can be made. The training can be "Formal Schools," conducted in-house, in a continuing education program at a local educational institution, or by consultants or vendors, or it can be directed "On-the-Job Training," or "Correspondence Courses"/Programmed Instruction materials. Whatever the choice, the bases for the "Training Programs" have been set forth by the analysts' and training specialists' efforts.

This ability to "Enhance" the individual's job value allows for "Career Development" by varying the individual's assignments within a given position category, by "Reclassification," when such is warranted due to increased job knowledge, and by "Promotion" when job responsibilities are increased significantly. This "Enhanced Job Value" for the individual is a form of job redesign; the "Enhanced Employment Value" provides the organization a more productive element with which more efficient and effective production can be achieved. Awareness of this increased productivity can be conveyed both to the individuals and to top level management by the use of the task-based "Performance Evaluation Review" form with an attendant feedback mechanism.

SUMMARY

As can be seen, by the process described above, data was developed for use in all of the aspects of human resources management. More specifically, however, it provided for: (1) "appraisal devices which were (are) actually job performance oriented"; (2) "through a thorough job analysis" specific KSA's for use in a "promotional potential" form which would be valid and the time involved in preparation is not prohibitive; (3) a choice among the alternatives available; (4) a performance appraisal device capable of being empirically validated as required by EEOC; and (5) they meet the specific needs and objectives of this organization. Further, employee participation was included in the initial data input; line management was involved in the development process; objective targets are being agreed upon and the data derived will be quantifiable; and finally, the managers are being trained in the purposes, procedures, interviewing skills, and feedback processes. Utilization of these forms has just begun. The initial reaction of both employees and supervisors appears to be favorable. However, since a semi-annual marking cycle is employed, it will be this time next year before statistically valid information can be provided.

APPLIED JOB ANALYSIS



BIBLIOGRAPHY

1. W. A. Evans, "Pay for Performance: Fact or Fable," Personnel Journal, September, 1970.
2. M. Schliemann, "The Use of Departmental Ratings in the Promotional Exam Process," Employment Division, Career Service Authority, Denver, Colorado, January, 1976.
3. W. M. Holley, H. S. Field, N. J. Barnett, "Analyzing Performance Appraisal Systems: An Empirical Study," Personnel Journal, September, 1976.
4. Ibid.
5. N. B. Winstanley, "The Use of Performance Appraisal in Compensation Administration," The Conference Board Record, March, 1975.
6. R. Beckett, B. Dodsworth, "Performance Appraisal -- Failed or Neglected," London, England, 1976.

RATIO OF FORCES TECHNIQUES FOR SCORING UNITS IN COMBAT SIMULATIONS

by

Dr. Edgar L. Shriver[1]
Kinton, Inc.

This paper concerns the scoring of two sided engagement simulations, as they have been developed in the U. S. Army. This is an on-going Army program which began as contract research by the Army Research Institute in 1972. Since that time, the methods for simulation and scoring have undergone several variations, and is the subject of continuing research, even as some elements are receiving world wide implementation. It is recognized that it is an open question as to how small unit combat simulation may best be scored. Others have dealt with the question during this session. It is a question which this paper will explore in some detail.

This paper does not represent official Army policy. The ideas are those of the author. The purpose of this paper is to present certain rationales to an interested, impartial audience with experience and expertise in military testing and to receive feedback on the theoretical points and rationales presented here today.

Now before going into the measurement problems, you need a thumbnail sketch of what two sided engagement simulation is. The simplest case is for small Infantry units, squad and platoons. Each soldier has a number on the front, rear and sides of his helmet. The size of the number is such that it can just be read at about 250 meters with a low power optical sight mounted on the rifle. At closer ranges it is easier to read, until at 50 meters it can be read with the naked eye. Now consider a soldier firing his rifle at ranges of 50 to

[1]Dr. Robert T. Root has been the Contracting Officer's Technical Representative (COTR) of engagement simulation research conducted for the Army Research Institute, and his thoughts contributed to the preparation of this paper.

250 meters. The probability of a hit is less and less just as is the probability of reading the number correctly is less and less. In fact, correlations of hit probability and number reading at different ranges is high. This provides a means for the soldiers of one unit to deliver the effects of rifle fire on another unit in close to real time in a two-sided simulated engagement.

There are other parts to this system; they include radios, and controllers to tap soldiers on the head when their number is called and to inform them that they are a casualty and that they should take their helmets off and lie down. There are similar systems for delivering the effects of all Infantry and Armor weapons including artillery fire. Without going into the details of how all this is done, the end product is a simulated battle between units where the ways in which leaders and troops behave determines what their effects will be on the other side - in approximately real time. These engagements are repeated several times a day with each side getting more effective with each "run". No two engagements are the same even if the units are the same and they start from the same points. Motivation is high in these engagements. The participants believe that the simulated engagements are like real combat and they learn fast to change the things that got them "killed" and to perfect what helped them produce casualties on the other side.

That sketch of engagement simulation is sufficient for our purposes here. We want to talk about measurement problems and ideas concerning how to create measurement systems to fit this situation.

There are inherent measurement problems in engagement simulation. In these contests, the two sides engage in a battle where each side delivers its

effects on the other until one or the other is so reduced by attrition that it is no longer capable of continuing the battle. It is not difficult to determine which side won or if the engagement is essentially a draw. But there are two general measurement problems: First, the individual contributions to the win or loss are not simple to measure. Second, win/lose measures do not take into account the relative strength of the two opponents.

For the contributions of each person to the unit outcome, there are these measurement problems:

1. By its nature each engagement is different. No individual encounters identical conditions in successive engagements, nor will his actions be identical. He learns to create new conditions to make it difficult for the other side.
2. Combat is a multivariate environment where a major objective is to prevent the other side from predicting future conditions from the conditions they first encounter.
3. Measurement situations traditionally depend on isolating variables, controlling conditions and predicting future performance on the basis of past performance.
4. The combat simulation cannot be controlled to meet traditional test conventions. To do so would be a classic case of the measuring instrument affecting what is measured.

It appears to us that measurement conventions will have to be created to fit the combat situation; so that the measuring instrument is not measuring the measuring instrument.

That's the individual measuring problem. Dr. Pulliam addressed that problem earlier this week. The present paper deals with the second problem, that of measuring the relative strengths of units with win/lose measures.

In measuring the unit, there is one datum which independent observers will agree upon. That is which side won. The simulated engagements continue until one side or the other is essentially destroyed; that is, until all personnel on the losing side are casualties or there are only a few lost survivors. If one side has zero personnel remaining and the other side has more, the assessment is merely a matter of counting. There is some ambiguity in distinguishing a tie from a win, e.g., zero survivors versus one survivor. But so long as the band of outcomes which represent a tie is sufficiently wide, this ambiguity is essentially resolved. So all units measured can be given a Won, Lost or Draw score. The win/lose measure reliably puts one side above the other. By "round robin" testing of all units this datum could be developed into an ordinal scale, i.e., A beat B, B beat C, etc. However, that amount of testing is impractical. We must explore other indexes of unit effectiveness.

In practice, one side may "win big". This means it destroys the other side with small loss to itself. The losses suffered in winning can then be used as an index of effectiveness. This too is a matter of counting survivors and is a reliable score. In addition to a score of Won or Lost, the percentage of the unit lost in accomplishing a Win can be reported, e.g., W 30%.

What kind of scale is this? Ordinal, Interval, Ratio? Let's see if it qualifies as a ratio scale. Would a loss of 10% be twice as good as a loss of 20%? I believe military commanders would find that approximately correct.

But they might also ask who was a casualty. All personnel in a unit are not equally effective. If the most effective soldiers were among the 10% lost, and the least effective among the 20%, the 20% loss might be preferable. In that case the scale would not even qualify as ordinal.

But assuming losses of random quality, the scale would be a ratio on this range of values. Now is a loss of 60% twice as bad as a loss of 30%? Perhaps, but one characteristic of this particular simulation is that the conflict is carried on beyond the point where a real combat engagement would probably be broken off. That means that a 60% loss is unlikely to occur in actual combat. That isn't to say that it hasn't happened. But real losses beyond 40% or 50% for one engagement are uncommon. A 60% loss is probably more than twice as serious as a 30% loss. Even so, we can consider the % casualties score to represent an interval scale, but not a ratio scale. That is, a 30% loss is better than a 31% loss, other things being equal, but a 60% loss is more than twice as bad as a 30% loss.

Let us explore another measure. First, the number of personnel on one "team" can be changed. This can't be done in a sport content; but in combat this is the rule rather than the exception. One combat tactic is to increase the size of your force in a particular area to the point where a breakthrough can be created. So changing the ratio of forces is the name of the game in combat. Our measurement system should be responsive to this characteristic. A unit that can win over superior odds is better than one that wins with equivalent odds. What kind of scale does this measure lead to? If unit A can defeat a force which is twice the size of a force that unit B can beat, is unit A twice as effective? If so, this is a ratio scale. In military terms, a commander can send one unit A to do the same job as two unit Bs. Either can win against the same opposing force. That certainly sounds like a ratio scale

That's fine, but still we have the practical problem of finding a homogeneously effective opposing force against which to make comparison. And if we find a standard criterion force, the process of conducting one engagement simulation to make the measurement of unit A may improve that criterion force by training, so that it is not the same when it engages unit B.

Quality may not change much in one engagement, so this method of measurement may be practical at certain ranges of quality. For instance, if unit A and unit B are both considerably higher in quality than a unit C, C may be used as a comparison force for both. That is, C will not increase in skill appreciably relative to A and B in one engagement. However, if A and B were only marginally better than C, the improvement in C due to one engagement would be a more important factor.

Setting aside this comparison technique for the moment, let us look for some baseline, or unity point unit.

There is a source of relatively homogeneous low level baseline personnel for measuring engagement simulation unit capability. A unit with no engagement simulation experience, recently formed of personnel newly out of Basic Training, might be considered to be of a homogeneous baseline quality. If unit A can win against twice the number of this baseline force that unit B can, then unit A is twice as good on a ratio scale. We can call them an X2 unit. That in turn should mean that a X4 unit should just about draw against two X2 units if this is a ratio scale.

To validate a X4 score obtained by winning against two X2 units, we would have to see if the X4 unit could win against four of the baseline units. This multiplication of factors might be validated in this way up through X4 or so. But at some point it becomes physically unrealistic to cram enough baseline

unit bodies together to test the ratio. The multiplier can go up indefinitely for more skilled units, but validating this ratio against baseline units becomes infeasible at some point. This problem would certainly become apparent in validating a X10 unit against 10 crammed up baseline units. Without validation, th ratio scale would "soften" into an interval scale. That is, we could no longer be sure that a X10 unit was worth two X5 units. But at any rate, we would have a ratio score up to some point and that it is at least an interval scale beyond that point. This ratio scale number is included with the win/lose score and the percentage lost score to form a three part score.

When units are approximately evenly matched, by quality times strength, the probability is that the winning side will indeed lose a fairly large proportion of its force. However, there is a possibility that they are "hardly scratched". A unit that has slight losses in winning against another unit that suffers heavy losses in losing, gets a lower % lost score but the same ratio of forces score. The score for a unit would include the win/loss component, the ratio of forces component and the % of force lost component, e.g., X4 W 20%. All indexes are meaningful in military terms. A possible way of integrating them will be discussed later. But there are some advantages to keeping the components separate.

The three part scoring system is like the old balance system of measuring weights. A few large weights can be placed in one measuring pan to balance a larger number of smaller weights in the other pan. The balance pans do not need to be loaded precisely - just well enough to get the ratio of forces in the right ball park. Beyond that, the % of personnel lost in achieving a win would add or subtract from the score of each side in the balance pans.

Units are placed in engagement simulations on the basis of the score they received on last measurement. After that, the losses they experience in achieving a win would add to the precision of the measurement. The object is first to get a rough balance of forces in the opposing pans. This may be accomplished with several units of small weight (weak forces) against a single one of larger weight (strong forces) or by forces of approximately equal weight. Depending on the outcome of the engagement their scores are adjusted up or down as a function of the % lost figure.

Having talked about measurement, let us turn to the psychology of the forces being measured. There are psychological purposes and motivations that the measurement system must be designed to support.

It is an important part of military psychology for units to learn that they can defeat a numerically superior force, especially a much larger force. Those skills are the same ones that enable them to expend fewer resources to defeat a smaller size force. But psychologically, winning over outside odds makes the members of a unit <u>feel</u> big and confident. Psychology is an important ingredient in battle. A unit that consistently wins over an over-size force in training will be psychologically prepared to do the same thing in combat. In actual combat, minimizing ones own losses is on each soldier's mind. Ther are means of avoiding personnel losses that are highly ineffective for tne unit, e.g., not firing. But beside avoiding personnel loss the qualities that enable units to win over odds are the same qualities that enable them to minimize their own losses. The three part score system is essentially measuring the same qualities on each part of the score.

There is still another psychological need that units have. They need to know that their tactics, plans and execution are as good or better than the other side's. If they have spent all their training in mopping up on a lot of weak forces, they won't be sure they can handle really good opponents in battle. The other side of this coin is that no unit likes to be beaten up by superior forces for too long. This implies that every unit should receive some experience against units of similar skill levels as well as against units of different skill levels. This implies loading the balance pans up in different ways, but always attempting to load them equally. How sure are we of our ability to load up both pans equally?

Using the score a unit achieved the last time they were measured seems an obvious way to load the balance pans. But it must be recognized that "a unit" has a turnover of personnel which probably changes its actual effectiveness, as compared to the score given a year or even six months previously. Keeping detailed historical records on turnover of personnel, and on how key they were to the unit's effectiveness, is a tedious approach. Let us consider what might happen if we allow units to adjust their previous score by a self-rating.

Let's say a unit has an old score of X2 W 20% and wants to self-rate itself as a X3 unit. It would then be joined by other units which had given themselves a similar self-rating. Let's say two such units are combined, to fight against an X6 self-rated unit. Let's say the X6 unit wins and expends half of its force in achieving the win, it would receive a rating of X6 W 50%. The losing units would receive a rating of X3 L 0%. How good is this rating in comparison to what they would have received if they had not adjusted their old score with a self-rating?

In other words, what the advantages and disadvantages of a unit self-rating itself high or low? A self rating that is intentionally too high subjects that unit to likely loss. Let's say a unit that is really X2 gives itself X4 rating under the rationale that they don't have anything to lose. They stack the odds against themselves so badly that they don't have much opportunity to win. It's not likely that they will "get lucky". But even if they lose they would retain the X4 rating accompanied with a L 0%. If they had correctly rated themselves at X2, they might have gotten an X2 W 50% rating indicating they had won. They won't be likely to get a W if they overrate themselves. But they could improve their ratio score by the self-rating. This raises a question of what component of the score to optimize on. Is it better to have an X2 W 50% score or a X4 L 0% score? The answer probably lies in what casualties were inflicted upon the other side. That is, if the X4 L 0% unit "hardly scratched" its opponents, it was overrated at X4 and less effective than an X2 W 50% unit. But, if the X4 L 0% unit inflicted more than 50% casualties on its opponent, it would be better than the X2 W 50% unit. This would imply still one more component to the scoring system: What the other side lost. That's a possibility we would like to explore at this point.

The losing units % lost score does not carry much information in the engagement simulation. That is, if the unit loses, its % lost score is going to be zero, or close to zero. The only reason for it not being zero is that one or more of its personnel were lost and couldn't be found. So for units that lose, we would provide more information by recording the losses they inflicted on the other side while going down to defeat. Let's consider what this does to our scoring system.

We had a X2 W 50% score and we wanted to compare it to a 4X L 0% score. We said the quality of the unit that lost depended on what losses it

on the winning unit. Let's say it inflicted 50% losses. We would now score it X4 L 50% instead of X4 L 0%. Now how does a X4 L 50% score compare to a X3 W 50% score? The losing unit inflicted 50% casualties on a force 4 times its size and lost its entire force. The winning unit lost 50% casualties in defeating a force twice its size but destroyed the entire losing force. Let's say each of the units being compared consisted of 10 men. The losing unit (X4) was against 40 men. They caused 50% casualties, or 20 men. They lost their entire force of 10 men.

The winning unit of ten men was against a force of 20 men. They destroyed those 20 and lost 5. In comparison both units destroyed 20 men. The losing unit lost 10 and the winning unit lost 5. At this point in the comparison the winning unit is better.

Without going into formulas, it is easy to see that losing units lose all their own personnel, and that's a big disadvantage. In short, it doesn't pay a unit to overrate themselves to the point where they make their defeat likely.

However, in the process of comparing a losing and winning unit we did two interesting things:

1. We substituted the loss inflicted on the winning side for the loss of the losing units own personnel and,
2. We combined the two scored components with each other to determine an overall figure of merit for comparing units with different force ratios, win/loss records and % losses.

We frankly don't know whether we have a ratio scale here, or an interval scale, when we combine components. It appears that there is at least an interval scale. It is clear that the score has a considerable ability to differentiate units. There are about 80 whole percentage numbers in each of two dimensions which can combine in all possible ways for 6400 possible scores. That is enough potential space to adequately spread all the Infantry squads in the Army.

Now let us return to the psychology of testing of units one more time. Remember we got into the question of self-rating because a unit will experience some turnover during a year, and will not be the same unit it was when it was tested the year before. Psychologically, a unit that achieves a certain ratio score, say X4, will probably work hard to keep that rating regardless of turnover of leaders or troops. That is, the psychological "unit identity" should create pressure on all new incumbents to "live up to" the unit score. That provides a rationale for keeping last year's score, and updating it with scores obtained during training engagements, rather than allowing self-rating. That's one option. But let us continue exploring the option of rating the unit by other means.

Let us suppose that senior commanders could update the ratio score for their units. In such a system, the commanders at higher levels would have "called their shots". They would receive feedback on these ratings they gave, after the tests were conducted. We would hypothesize that the results would be more "attention getting" to commanders that have factored in their judgements. The reason is that everyone is interested in feedback on their own predictions. Just being in the predictive process could make the results more meaningful to the commanders. The results might become a subject of informal discussion for

days or even weeks, and the identities of outstanding units would be passed around. The value of personal identification in commanders' eyes is psychologically more rewarding to personnel than almost any other form of reward. Losers can always get to work on "next time". But the big winners become personally known. What they did becomes the goal for everyone the next time. That kind of reward will do more to improve the utilization of training time than any set of check offs on schedule charts. When soldiers and units really believe that the test results give an accurate reflection of how good they will be in combat, and that their achievements are known to higher level commanders, this increases the value of the measurement system beyond mere assessment. Whether this process would take place regardless of whether the unit commanders factored in their judgements is not known. It is just an idea. A further idea on this point is to determine how well commanders know the capability of each of their own units. That is, it would not be sufficient to say "all my units are outstanding". The good commander would be the one that knows his units and can say how good they are.

There may be a question about why to bother with a three part evaluation score. Why not just put the units out there and let A go against B. Well, that method is OK at the beginning of engagement simulation training when all units will be at a common low level. But the objective is to use measurement to build motivation as well as assess units. The system of rewards should be a motivation. This may not be a traditional goal of testing, but we think it is not only legitimate but necessary as a testing goal. Differentiation of units is the key to motivation. The motivational leverage is in identifying great differences in unit's skill. A unit that can win over a force six times its size is not merely "another unit". It is a goal object. The objective is

:o show that such feats are possible. The system of balance pans is needed :o bring this out. Just sending any old A against B will not bring out the ›otential improvements that are possible. The measurement system which gives ı believable score on an interval or sometimes ratio scale is a necessary ingredient. This is different than the concept of bringing all units up to ;tandard performance. Psychologically, reaching equivalent performance is a ıull objective. Units may work to avoid being the "goat". But the objective ›f standard performance is to ignore excellence. Under this "standard level" :oncep', hairs must be split to find an "outstanding" unit. The proposed ıpproach is one of opening up "space" between unit levels of skill. It may :ake several years for units to fill in the space. But the measurement system ıust be in place at the beginning and have the potential for believably scoring ı space, with a high ceiling. Then it must show that it has the potential for ıeasuring that open space in terms that are believable to participants.

We are dealing with ideas for a measurement system that is credible, valid ınd motivating. Your views are solicited on any aspect of such a measuring ;ystem for engagement simulations.

LIST OF CONFEREES

1. ACKER, BG William P.
   Deputy Commander, Department of Defense, HDQ US Military Enlistment Processing Command, Fort Sheridan, IL 60037

2. ADAMS, William
   CNET, Naval Air Station, Pensacola, FL 32508

3. ALBECK, Robert D.
   Curriculum & Instructional Stds Office, NTTC Corry Station, Pensacola, FL 32508

4. ALLEY, William E.
   Air Force Human Resources Laboratory, Lackland Air Force Base, TX 78236

5. ANDERSON, CAPT, Erns M.
   Officer in Charge, Navy Occupational Development & Analysis Center, Bldg., 150, Washington Navy Yard (Anacostia), Washington, D.C. 20374

6. ANDERSON, Harry R.
   US Coast Guard Aviation Training Center, Bates Field, Mobile, AL 36608

7. ANSBRO, Thomas M.
   NAVEDTRAPRODEVCEN, CTAG, Pensacola, FL 32509

8. ANTHONY, Jack
   Office of Directorate of Evaluation, US Army Field Arty School, Fort Sill, OK 73503

9. ARNOLD, CWO Brian D.
   Department of Training Development, School of Music, USNAB (Little Creek), Norfolk, VA 23521

10. ASHBURN, LTCDR J.H.
    NAMRL, Pensacola, FL 32508

11. AYRES, 1LT Gregory D.
    Educational Test/Evaluation Branch, Asst Sup for Instruction, Academy of Health Sciences, Fort Sam Houston, TX 78234

12. BACHTEL, Mary A.
    NAVEDTRAPRODEVCEN, CODE PD-9, Pensacola, FL 32509

13. BAISDEN, Annette G.
    NAMRL, Pensacola, FL 32508

14. BALL, William C.
Director, Training Developments Directorate, US Army Ordnance Center & School, Aberdeen Proving Ground, MD 21005

15. BARANOWSKI, Philip J.
NAVEDTRAPRODEVCEN, CODE EA, Pensacola, FL 32509

16. BARRON, C.J.
NAVEDTRAPRODEVCEN, CODE PD-10, Pensacola, FL 32509

17. BEJAR, Isaac, I.
Dept of Psychology, N-660, Elliott Hall, University of Minnesota, Minneapolis, MN 55455

18. BEGLAND, CPT Bob
Florida State University, Center for Ed Tech, Tallahassee, FL 32303

19. BERNSTEIN, D.
Career Patterns and Training Programs Branch, Reserve Training Div., Office of Reserve, Coast Guard Headquarters, Washington, DC 20590

20. BIRDSALL, Walter W.
NAVEDTRAPRODEVCEN, CODE PD, Pensacola, FL 32509

21. BLAND, LCDR R.D.
US Coast Guard Institute, PO Sub-Station 18, Oklahoma City, OK 73169

22. BLEILER, Paul
NAVEDTRAPRODEVCEN, CODE PD-4, Pensacola, FL 32509

23. BLUMENFELD, DR Warren S.
Georgia State University, University Plaza, Atlanta, GA 30303

24. BOI, ENSIGN K.J.
Coast Guard Headquarters, Training & Education, Washington, DC 20590

25. BORDASS, LTC Timothy D.
Training Development Directorate, US Army Ordnance Center & School, Aberdeen Proving Ground, MD 21005

26. BOUZA, Duane J.
USAF Base Education Center, (DPT) Offutt Air Force Base, NB 68113

27. BOYD, DR Joseph L.
Educational Testing Service, Princeton, NJ 08540

28. BRAUN, Frank B.
Data-Design Laboratories, 15 Koger Executive Center, Norfolk, VA 23501

29. BRITTAIN, DR Clay V.
Army Individual Training Evaluation Group, Fort Eustis, VA 23604

30. BROGA, LCDR Dwight C., III
USCG HQ, 400 Seventh St., S.W., Washington, DC 20590

31. BROWN, Victor H.
NAVEDTRAPRODEVCEN, CTAG, Pensacola, FL 32509

32. BUCK, DR C.W.
USCG Institute, PO Sub-Station 18, Oklahoma City, OK 73169

33. BURT, John A.
USCG Institute, PO Sub-Station 18, Oklahoma City, OK 73169

34. CALLAHAN, M.D.
Head, NOTAP Dept, NODAC, Bldg. 150, WNY (Anacostia), Washington, DC 20374

35. CARPENTER, LCOL G.J.
Royal Military College of Canada, 3 Ridout Row, RMC, Kingston, ONT K7L2W3

36. CARPENTER, DR James B.
St Mary's University, One Camino Santa Maria, San Antonio, TX 78284

37. CARRAWAY, Jay
NAVEDTRAPRODEVCEN CODE PD-6, Pensacola, FL 32509

38. CHIPPENDALE, Joan T.
USCG Training Center, Governors Island, NY 10004

39. CHRISTAL, DR Raymond E.
AFHRL/OR USAF, Lackland Air Force Base, San Antonio, TX 78236

40. COCKRELL, DR John T.
Army Research Institute, Fort Knox Field Unit, Fort Knox, KY 40121

41. COLEMAN, Don
NTTC Corry Station CODE 01E, Pensacola, FL 32511

42. COREY, Lawrence G.
Plog Research Inc., 15840 Ventura Blvd., Encino, CA 91436

43. CORMACK, CDR Bruce (Ret'd)
Bruce Cormack Associates Limited, 24 Glenridge Road, Ottawa, Ontario, Canada K2G 2Z5

44. CORY, DR Charles H.
Navy Personnel Research & Development Center, San Diego, CA 92152

45. COTHRON, Joe M.
HQ US Army, TAGO/TAGCEN, Education Directorate, Forrestal Building, Independence Avenue, Washington, DC 20314

46. DAVIS, DR Charles E.
ONR-Branch Office Chicago, 536 South Clark Street, Chicago, IL 60605

47. DAVIS, D.D.
Chief of Naval Education and Training, Naval Air Station, Pensacola, FL 32508

48. DAVIS, MAJ Melvin L.
Directorate of Training Developments, US Army Aviation Center, Fort Rucker, AL 36362

49. DAVIS, Thomas
Office of Manpower Utilization, HQMC (MPU), Quantico, VA 22134

50. DICKINSON, R.W.

51. DOLL, Richard
102 Bear Drive, Gulf Breeze, FL 32561

52. DOORLEY, Richard D.
Evaluation Systems Office, USA MILPERCEN, 200 Stovall Street, RM 1523, Alexandria, VA 22332

53. DUBOSKY, Stephen W.
Director of Training Development, USAAVNC, Fort Rucker, AL 36362

54. DUFFY, Paul C.
Marine Corps Institute, Marine Barracks, 8th & I Sts., S.E. Washington, DC 20390

55. DUTTON, DR Maurice G.
President, ARBEC, Inc., 3909 G North Interregional, Austin, TX 78722

56. EASSON, Pamela J.
DPCA Army Education Center, Hunter Army Airfield, GA 31409

57. EASTMAN, R.F.
ARI Field Unit, PO Box 476, Fort Rucker, AL 36360

58. EISENTROUT, CAPT Bruce M.
Directorate of Evaluation, USAARMS, Fort Knox, KY 40121

59. EISMAN, Ed
Navy Instructional Technology Development Center, Bldg. 499, Code 7000, USNTC, San Diego, CA 92133

60. FOLEY, Paul
Navy Personnel Research & Development Center, San Diego, CA 92152

61. FREY, Virgil E.
NAVEDTRAPRODEVCEN, CTAG, Pensacola, FL 32509

62. FROM, E.E.
NAVEDTRAPRODEVCEN, CODE EA, Pensacola, FL 32509

63. GALLAGHER, CAPT J.T.
Naval Aerospace Medical Research Lab, NAS, Pensacola, FL 32508

64. GALLAWA, Jacob
NAVEDTRAPRODEVCEN, Corry Station, Code PD-2, Pensacola, FL 32511

65. GARY, Perry L.
DTD (ATZQ-TD-TAD-TD), Fort Rucker, AL 36301

66. GEESEY, Scott S., Jr.
Individual Training Design Section, Directorate of Training Development, USA Quartermaster School, ATSM-TD-IT, (Bldg. T-4201), Fort Lee, VA 23801

67. GERBER, DR J.E., JR.
US Army Forces Command (AFPR-PSE), Fort McPherson, GA 30330

68. GIBSON, Rodney H.
Course Development Dept/Institutional Training Branch, Fort Rucker, AL 36362

69. GILBERT, DR A.C.F.
US Army Research Institute, 1300 Wilson Blvd., Arlington, VA 22209

70. GIORGIA, M. Joyce
AFHRL/ORS, Lackland Air Force Base, TX 78236

71. GLESZER, MAJOR Peter E.
Staff & Faculty Training Division, USAIS, Fort Benning, GA 31905

72. GOODGAME, Doug
Occupational Research, Industrial Engineering Dept., Texas A&M University, College Station, TX 77843

73. GOODY, SQNLDR K.
Royal Australian Air Force, AFHRL/OR Stop 63, Lackland Air Force Base, TX 78236

74. GORDON, M. Meriwether
AFROTC/ACME, Education Evaluation, Maxwell Air Force Base, AL 36112

75. GORMAN, Steven
HQ Marine Corps, Washington, DC 20380

76. GOULD, R. Bruce
AFHRL/PES, Lackland Air Force Base, TX 78209

77. GRANT, R.F.
NAVEDTRAPRODEVCEN, Corry Station, Pensacola, FL 32511

78. GRIMM, Richard L.
NAVEDTRAPRODEVCEN, CTAG, Pensacola, FL 32509

79. GUY, Robert L.
NAVEDTRAPRODEVCEN, CODE PD-3, Pensacola, FL 32509

80. HALTRECHT, DR Ed
Personnel Research, H2-D17, 700 University Avenue, HYDRO, Toronto, Ontario

81. HAMELOTH, Wayne H.
CISO Corry Station, Pensacola, FL 32511

82. HAMILTON, Charles
Air Force Military Personnel Center, Research Division, Randolph Air Force Base, TX 78148

83. HANSON, CPT Verne B.
Evaluation Division, US Army Administration Center, IG/C,E Fort Benjamin Harrison, IN 46216

84. HARMAN, Samuel A.
TNG. DIV.-USAFESA, Fort Belvior, VA 22060

85. HAWKINS, DR Helmut H.
USA-MILPERCEN, 2461 Eisenhower Avenue, Alexandria, VA 22331

86. HAYES, William A.
NAVEDTRAPRODEVCEN, CTAG, Pensacola, Florida 32509

87. HEAD, Lawrence W.
USAF School of Applied Cryptologic Sciences, Goodfellow Air Force Base, Texas 76901

88. HERMANSEN, Art
ITEG, Fort Eustis, VA 23604

89. HICKS, DR Jack M.
Army Research Institute, 1300 Wilson Blvd, Arlington, VA 22209

90. HODGES, Will
Army Education Center, Fort Stewart, GA 31313

91. HOGGATT, COL Ralph S.
Air Force Human Resources Laboratory, Brooks Air Force Base, TX 78235

92. HOGUE, Ken
Occupational Research, Industrial Engineering Department, Texas A&M University, College Station, Texas 77843

93. HOLGATE, CPT Stanley H.
BIOMED Lab, Aberdeen Proving Ground, MD 21010

94. HOLLIER, I.S.
Office of Manpower Utilization HQ US Marine Corps, Code MPU, Quantico, VA 22134

95. HORAN, LT William T.
Commandant (G-PMR) US Coast Guard, Washington, D.C. 20590

96. HOSMER, Clark
39 Longwood Drive, Shalimar, Florida 32579

97. HOUTZ, John C.
US Army Recruiting Command, Fort Sheridan, Illinois 60037

98. HOYT, DR William G.
System Development Corporation, 2500 Colorado Ave., Santa Monica, California 90406

99. HUDSON, CWO3 R.M.
Coast Guard Leadership Programs Training and Development Staff, Reserve Training Center, Yorktown, VA 23690

100. JACKSON, William L.
Directorate of Standardization, USAAVNC, Fort Rucker, AL 36362

101. JEFFRIES, CDR W. H.
Naval Technical Training Center, NTTC Corry Station, CODE-50, Pensacola, FL 32511

102. JOAQUIN, CAPTAIN John
Canadian Forces Personnel Applied Research Unit, 1107 Avenue Road, Toronto, Ontario M5N 2E4

103. JOHNSON, Cecil D.
US Army Research Institute for the Behavioral & Social Sciences, 1300 Wilson Blvd., Commonwealth Bldg., Room 241, Arlington, VA 22209

104. JOHNSON, Dorothy J.
Naval Technical Training Center, CODE (01E), Corry Station, Pensacola, FL 32511

105. JOHNSON, Eugene, III
US Army Research Institute for the Behavioral & Social Sciences, 1300 Wilson Blvd., Commonwealth Bldg., Arlington, VA 22209

106. JOHNSON, Robert N.
US Army Administration Center, Training Dev Directorate, Fort Benjamin Harrison, Indianapolis, IN 46216

107. JONES, Karen N.
Test Development, Coast Guard Institute, PO Sub-Station 18, Oklahoma City OK 73169

108. KANTOR, DR Jeffrey E.
Air Force Human Resources Laboratory, AFHRL/PEM, Stop 63, Lackland Air Force Base, San Antonio, TX 78236

109. KAUFFMAN, Britt
University of Texas at Austin, 4901 Ave F, Austin, TX 78751

110. KENNEDY, LT COL Charles R.
USAF Occupational Measurement Center, Lackland Air Force Base, TX 78236

111. KETTLEHUT, CAPT Allan J.
Squadron Officer School, Maxwell Air Force Base, AL 36113

112. KNERR, DR Bruce W.
US Army Research Institute, 1300 Wilson Boulevard, Arlington, VA 22209

113. KNERR, DR Claramae S.
US Army Research Institute, 1300 Wilson Boulevard, Arlington, VA 22209

114. KORB, Lawrence J.
Professor of Management, US Naval War College, Newport, R.I. 02840

115. KOROTKIN, DR Arthur L.
Richard A. Gibboney Associates, Inc., 10605 Concord St., Suite 203A, Kensington, MD 20795

116. KUENZ, DR Marjorie A.
Naval Health Sciences Education and Training Command, HSETC-21, N.N.M.C., Bethesda, MD 20014

117. LANG, Darryl
US Coast Guard, Washington, DC 20590

118. LANTERMAN, Richard S.
HQ, US Coast Guard, G-P-1/2/62, Washington, DC 20590

119. LAUMEYER, Norman G.
Extension Training Development Branch, Directorate of Training Development, Fort Rucker, AL 36362

120. LESTER, James
Office of Naval Research, Boston Branch Office, ONR - 495 Summer Street, Boston, MA 02210

121. LEWISKI, DR Raymond
ECI/EDV, Gunter Air Force Station, AL 36118

122. LIBERTY, DR Paul G.
Vice President, Research and Evaluation, ARBEC, INC., 3909 North Interregional, Austin, TX 78722

123. LUPONE, D.F.
NAVEDTRAPRODEVCEN, CODE PD-5, Pensacola, FL 32509

124. MASCIANGELO, MAJ William R.
Office of Manpower Utilization, Quantico, VA 22134

125. MAY, DR William E.
Special Assistant to the Commandant for Faculty Education, US Army Missile and Munitions Center & School, Redstone Arsenal, AL 35809

126. McAREAVY, DR John F.
Dept. of Public Administration, US Army Management Engineering Training Activity, Rock Island, IL 61201

127. McBRIDE, James R.
Army Research Institute, 1300 Wilson Boulevard, Arlington, VA 22209

128. McCABE, Jerome M.
Data-Design Laboratories, 1755 South Jefferson Davis Highway, Suite 307, Arlington, VA 22202

129. McCLINTOCK, DR W.R.
NAVEDTRAPRODEVCEN, CODE PD-10, Pensacola, FL 32509

130. McFARLAND, CAPT Barry P.
USAF ASD/ENECC, Wright-Patterson Air Force Base, OH 45424

131. McFARLING, CPT Les
Academy of Health Sciences, HSA-ITE, Fort Sam Houston, TX 78234

132. McINTOSH, Vergil M.
Extension Course Institute, Air University, ECI/EDV, Gunter Air Force Station, AL 36118

133. McMULLEN Robert L.
Army Research Institute Field Unit, Box 476, Fort Rucker, AL 36362

134. McVAY, K

135. MEREDITH, John B., Jr.
Data-Design Laboratories, 15 Koger Executive Center, Suite 140, PO 12773, Norfolk, VA 23502

136. MESSICK, Vernon D.
NAVEDTRAPRODEVCEN, CODE PD-9, Pensacola, FL 32509

137. MIGUEL, DR Richard
Center for Vocational Education, Ohio State University, 1960 Kenny Road, Columbus, OH 43210

138. MILATZO, John P.
National Security Agency, 9800 Savage Road, Fort Meade, MD 20755

139. MILLER, CDR James P.
Naval Personnel Program Support Activity, Bldg. 150, Naval Station Anacostia, Washington, DC 20374

140. MIRABELLA, DR Angelo
Army Research Institute, 1300 Wilson Boulevard, Arlington, VA 22209

141. MOBLEY, Amelia E.
US Coast Guard, 400 7th Street, S.W., Washington, DC 20590

142. MOONAN, DR William J.
Navy Personnel Research & Development Center, CODE 9005, San Diego, CA 92152

143. MORGAN, DR Robert M.
Instructional Systems Development Center, 206 Dodd Hall, Florida State University, Tallahassee, FL 32306

144. MURPHY, John W.
USA IA Evaluation Division, Fort Benjamin Harrison, Indianapolis, IN 46226

145. NEFF, Edwin F.
US Coast Guard Headquarters, G-PTE/72, 400 Seventh St, S.W., Washington, DC 20590

146. NELSON, Ray W.
Naval Technical Training Center (OIE), NTTC Corry Station, Pensacola, FL 32511

147. NOVAK, Frank J.
NAVEDTRAPRODEVCEN, CODE PD-9, Pensacola, FL 32509

148. O'BRIEN, R.
Test & Measurement Spec, DLISDA-ELD, Lackland Air Force Base, TX 78236

149. OCCHIALINI, Chris
PACDA-HRD, Building 1, Fort Benjamin Harrison, IN 46249

150. O'NEIL, Harold F.
Advance Research Projects Agency, 1400 Wilson Blvd., Arlington, VA 22209

151. ORGANT, Morgan
NAVEDTRAPRODEVCEN, CODE PD-8, Pensacola, FL 32509

152. OSBORN, William C.
HUMRRO, 632 Knox Blvd., Radcliffe, KY 40162

153. OSBORNE, James E.
NAVEDTRAPRODEVCEN, CODE PD-3, Pensacola, FL 32509

154. OWEN, D.

155. PATRICK, W.T.
NAVEDTRAPRODEVCEN, CODE PD-7, Pensacola, FL 32509

156. PINE, DR Steven M.
N660 Elliott Hall, University of Minnesota, Minneapolis, MN 55455

157. PIPKIN, C.M.
NAVEDTRAPRODEVCEN, CODE PD-7, Pensacola, FL 32509

158. PLAUT, MAJOR Peter K.
The Judge Advocate General's School, Charlottesville, VA 22901

159. PORTER, George V.
Director, Cadet Examinations & Records, USAF Academy, CO 80840

160. PRATT, William R.
NAVEDTRAPRODEVCEN, CODE PD-6, Pensacola, FL 32509

161. PRITCHARD, Ronald J.
Directorate of Evaluation, Fort Knox, KY 40121

162. PROPHET, Wallace W.
Human Resources Research Organization 400 Plaza Building, Pensacola, FL 32505

163. PULLIAM, DR Robert
10242 Stratford Avenue, Fairfax, VA 22030

164. RAFACZ, Bernard A.
Navy Personnel R & D Center, San Diego, CA 92152

165. RASTALL, DR G.D.
CNET, CODE N-101, Naval Air Station, Pensacola, FL 32508

166. RAYNER, DR Gail T.
Florida State University, 1A Tully Gym, Tallahassee, FL 32306

167. REE, DR Malcolm
Air Force Human Resources Lab., AFHRL/PES, STOP 63, Lackland Air Force Base, TX 78236

168. RODEEN, COL
MEPCOM, Fort Sheridon, IL 60037

169. ROSE, DR Andrew M.
American Institutes for Research, 1055 Thomas Jefferson Street, N.W. Washington, DC 20007

170. ROUGHGARDEN, CDR K.M.
USCG Institute, PO Sub-STATION 18, Oklahoma City, OK 73169

171. RUSH, John R.
NAVEDTRAPRODEVCEN CODE PD-6, Pensacola, FL 32509

172. SANDS, DR Franklin F.
Navy Personnel Research & Development Center, CODE 201, NPRDC, San Diego, CA 92152

173. SANDS, William A.
Navy Personnel Research & Development Center, San Diego, CA 92152

174. SCANLAND, DR Francis W.
Assistant Chief of Staff for Research and Program Development, Chief of Naval Education and Training, Naval Air Station, Pensacola, FL 32508

175. SCHRAM, Richard J.
Curriculum & Instructional Standards Ofc., CODE 01E, NTTC, Corry Station, Pensacola, FL 32511

176. SCHOFIELD, D.L.

177. SEIBEL, D.

178. SELLMAN, MAJOR Wayne S.
Air Force Military Personnel Center, AFMPC/DPMYO, Randolph Air Force Base, TX 78148

179. SEUBERLICH, LT COL H.E.
Chairman Army Service, (DBWV), 5300 Bonn-Bad, Godesberg, Sudstrabe 123, Germany

180. SHIPLEY, B.D.
US Army Research Institute for the Behavioral & Social Sciences, PO Box 476, Fort Rucker, AL 36362

181. SHOEN, W.R.
Senior Educational Advisor, Service School Command, Naval Training Center, Orlando, FL 32813

182. SHRIVER, DR E.L.
7245 Arlington Blvd., Matrix Research Co., Falls Church, VA 22042

183. SHULL, MAJOR Warren B.
Air Force Reserve Officers' Training Corps, AFROTC/SDEC, Maxwell Air Force Base, AL 36112

184. SIMMONS, C. Ray
Data-Design Laboratory, 7925 Center Avenue, Cucamonga, CA 97130

185. SINGLETARY, James W.
NAVEDTRAPRODEVCEN, CTAG, Pensacola, FL 32509

186. SMITH, DR M.J.
NAVEDTRAPRODEVCEN, CODE EA, Pensacola, FL 32509

187. SMITH, Reginald
Naval Technical Training Center, CODE (01E) Corry Station, Pensacola, FL 32511

188. SNYDER, DR Conrad Wesley, Jr.
American Institutes for Research, 1055 Thomas Jefferson St., N.W., Washington, DC 20007

189. SNYDER, CH (LTC) John S.
Directorate of Evaluation, US Army Chaplain Center & School, Fort Wadsworth, Staten Island, NY 10305

190. SOLOMON, E.G.
NAVEDTRAPRODEVCEN, CODE PD-5, Pensacola, FL 32509

191. SPECTOR, DR Daniel E.
US Army Military Police School, ATSJ-DTD-AD, Fort McClellan, AL 36205

192. STEARNS, Ralph I.
USAF, Base Education Center (3902-ABWg/DPT), Offutt Air Force Base, NB 68113

193. STEIN, CPT Earl S.
US Army Natick Research & Development Command, Natick, MA 01760

194. STEINHEISER, DR Frederick
Army Research Institute, 1300 Wilson Blvd., Arlington, VA 22209

195. STENGEL, LTCOL
MEPCOM, Fort Sheridan, IL 60037

196. STONE, CAPT B.G.
Chief of Naval Education and Training, Naval Air Station, Pensacola, FL 32509

197. SULLIVAN, Denis J., Jr.
US Army Intelligence School, Fort Devens, MA 01433

198. SWANSON, SSG Peter J.
Dept of Tng Dev., School of Music, USNAB (Little Creek) Norfolk, VA 23521

199. SWANSON, MAJOR Ronald G.
Academic Instructor School, AIAOS/EDV, Maxwell Air Force Base, AL 36112

200. SYMINGTON, DR Lawrence E.
US Army Natick Research & Development Command, Natick, MA 01760

201. SYMPSON, James B.
Research Assistant, N650 Elliott Hall/Dept of Psychology, University of Minnesota, Minneapolis, MN 55455

202. TAKAHASKI, Terry Y.
Defense Intelligence School, Washington, DC 20374

203. TARTELL, CAPT J.S.
USAF Occupational Measurement Center, Lackland Air Force Base, TX 78236

204. TAYLOR, CDR Michael O.
Officer in Charge, Naval Manpower Utilization Unit, HMS Vernon, Portsmouth, Hampshire, England

205. TAYLOR, Stephen R.
Directorate of Training Developments, US Army Infantry School, Fort Benning, GA 31905

206. THOMAS, CAPT R.T.
Commanding Officer, NAVEDTRAPRODEVCEN, CODE-A, Pensacola, FL 32509

207. TINDALL, J.E.
Data-Design Laboratories, 15 Koger Executive Center, Norfolk, VA 23502

208. TORMEY, Paul D.
US Army Logistics Management Center, Fort Lee, VA 23801

209. TRACY, John M.

210. TUCKER, LT Charles W.
Chief of Naval Personnel, (PERS-551), Washington, DC 20370

211. TURLEY, LCDR L.A.
Royal Australian Navy, 401 N 77th Ave., Pensacola, FL 32506

212. UHLANER, DR J.E.
US Army Research Institute for the Behavioral & Social Sciences, 1300 Wilson Blvd., Rossyln, VA 22209

213. VALE, C.D.
Dept of Psychology, N660 Elliott Hall, University of Minnesota, Minneapolis, MN 55455

214. VALENTINE, DR Lonnie D., Jr.
Air Force Human Resources Laboratory, Stop 63
Lackland Air Force Base, TX 78236

215. VAN CLEVE, DR Roy R.
Comptroller of Public Accounts, State of Texas,
LBJ Bldg., 17th & Brazos, Austin TX 79167

216. VAUGHAN, CAPT David S.
USAF Occupational Measurement Center, Lackland Air
Force Base, TX 78236

217. WALDKOETTER, Raymond
US Army Research Institute, 1300 Wilson Boulevard,
Arlington, VA 22209

218. WALKER, LCDR C.L.
Central Test Site for PTEP, Naval Guided Missile
School, Dam Neck, VA 23461

219. WALSH, Frank B.
USA Intelligence School, Fort Devens, MA 01433

220. WASER, Charles
US Coast Guard, US Coast Guard Headquarters,
(G-P-1/62), Washington, DC 20590

221. WEISSMULLER, Johnny J.
AFHRL/SMAZ Stop 63, Lackland Air Force Base, TX 78236

222. WIANT, Allen A.
Center for Vocational Education, Ohio State University,
1960 Kenny Road, Columbus, OH 43210

223. WILFONG, DR Harry D.
MEPCOM, Fort Sheridan, IL 60037

224. WILLING, Richard C.
US Coast Guard Institute, PO Sub-Station 18, (MVP),
Oklahoma City, OK 73169

225. WILSON, VADM J.B.
Chief of Naval Education and Training, Bldg. 624,
Naval Air Station, Pensacola, FL 32508

226. WILTSHIRE, CPT Robert B.
Individual Training Evaluation Group, Fort Eustis,
VA 23604

227. WINIEWICZ, Casimer
Staffing Division, Chicago Region, 230 S. Dearborn St.
Chicago, IL 60640

228. WORSTINE, Darrell A.
US Army Military Personnel Center, 2461 Eisenhower Ave.
Alexandria, VA 22331

229. WYPICK, CDR John M.
USCG Training Center, Government Island,
Alameda, CA 94501

230. YAMAMOTO, CPT Richard Y.
Directorate of Evaluations, US Army Military Police
School, USAMPS, Fort McClellan, AL 36205

231. YARUSSO, Lowell
NAVEDTRASUPPCENLANT, Bldg. Z-86, Naval Station,
Norfolk, VA 23511

232. YELLEN, Ted
Navy Personnel Research & Development Center,
San Diego, CA 92152

233. YORK, William J., Jr.
US Army Signal School, ATSN-DTD-IT-T, Fort Gordon,
GA 30905

234. ZAGORSKI, COL F.J.
Defense Intelligence School, Washington, DC 20374

235. ZELLER, DR Anchard F.
Air Force Inspection & Safety Center, AFISC/SEL,
Norton Air Force Base, CA 92409

BY-LAWS OF THE MILITARY TESTING ASSOCIATION

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems. Each agency shall have no more than two (2) professional civilian representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. A majority of the members of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the association, and the Steering Committee, and to conduct official correspondence of the association, and to insure notices for conferences. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment in order of the following rotating schedule:

United States Army
United States Marine Corps
United States Navy
United States Air Force
United States Coast Guard

C. The annual conference of the Association shall be held at a time and placed determined by the coordinating agency. The membership of the association shall be informed at the annual conference of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the Chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be an ex officio member of all standing committees.

B. The President with the counsel and approval of the Steering Committee may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in consultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-third's vote of the assembled membership of the association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Enactment

These By-Laws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 2 November 1973).

REPORT OF STEERING COMMITTEE
AND
GENERAL BUSINESS MEETINGS

1. NAME OF THE ORGANIZATION

1363 questionnaires were sent out requesting suggestions for the name of the organization. Only 58 replies were received. There were no trends or similarities in these replies. Both the Steering Committee and the General Assembly voted to retain the current name, "Military Testing Association."

2. U. S. CIVIL SERVICE COMMISSION

Because of a communication problem, it was reported that the U. S. Civil Service Commission had declined the membership invitation of the 1975 meeting. This was erroneous. The commission "...welcomes the invitation for Associate Membership in MTA."

3. ARTICLE III OF THE BYLAWS

The definitions of primary and associate membership were discussed. The matter was tabled to be brought up at the 1977 meeting.

4. ARTICLE VII OF THE BYLAWS

Section B of the Bylaws deals with the hosting responsibilities of the annual conference. As there is now more than one agency within a given service, rotating hosting responsibilities must be established. Amending this Article was delayed until Article III is amended. However, hosting sites were established for the next four years.

1977	San Antonio - Air Force Human Relations Laboratory
1978	Oklahoma City - Coast Guard Institute
1979	San Diego - Navy Personnel Research and Development Center
1980	Toronto, Canada - Canadian Forces Personnel. Applied Research Unit

INDEX

Acker, William P.
The Joint Services Approach to Enlistment Processing.....620

Alley, William E.
Vocational Interest Assessment in the Air Force.....984

Anderson, Harry R. and Povenmire, H. Kingsley
Total Testing in Coast Guard Aviation Training.....193

Bain, Captain Earl H. and Joaquin, Captain John B.
Evaluation of Junior Leadership Assessment Procedures.....1197

Baisden, Annette G.
Research on the Aeronautical Adaptability of Women.....541

Bejar, Isaac I.
Applications of Adaptive Testing in Measuring Achievement and Performance.....504

Bland, Raymond D., LCDR
Indirect Sampling of the Affective Domain To Assess "Petty Officer Qualities".....142

Blumenfeld, Warren S.
Efficacies of Three Methods of Weighting Biodata to Predict Turnover.....425

Boycan, G. Gary and Rose, Andrew M.
Evaluation of Crew Gunnery Performance.....1184

Brittain, Clay V.
Establishing the Validity of SQT.....292
Introduction to the Symposium.....270

Christal, Raymond E.
What is the Value of Aptitude Tests?.....36

Cockrell, John
Television as Stimulus Input in Synthetic Performance Testing Experiment I.....218

Cory, Charles H.
The Predictive Validity of Operational and Experimental Variables for Mental Group IVS in the Navy: A Review and Summary of the Findings from Four Sets of "Culture Fair" Tests.....812

Corey, Lawrence G.
A Segmentation Analysis of Women in the Army.....551

Davis, D.D. and Ansbro, T.M.
Occupational Analysis for Navy Instructional Systems Development (ISD): A Matrix Approach.....662

Dutton, Maurice, Ph.D. and Liberty, Paul, Ph.D.
The Evaluation-in-Systems (EIS) Approach in Development of Training Programs.....1117

Eastman, Robert F. and McMullen, Robert L.
Reliability of Associate Ratings of Performance Potential by Army Aviators.....182

Epstein, Kenneth I.
Performance Objectives: Problems in Definition Measurement, and Interpretation.....1151

Erwin, Donald E.
Evaluative Criteria for Comparing Models of Unit Performance.....1070

Gerber, J. E. Jr., Ph.D.
"The Other Ten Percent" - Revisited.....353

Gilbert, Arthur C. F., Ph.D. and Grafton, Frances C.
Characteristics of an Officer Performance Evaluation Measure.....1136

Gorman, Steven
The Utilization of Women in the Marine Corps.....530

Gould, R. Bruce
An Air Force Job Satisfaction Research Project through 1976.....368

Griffin, John C., DeBerg, Oak H. and McFarland, Barry P.
Evaluation of G-Seat Cueing in Large Bodied Aircraft.....904

Hermansen, Arthur G.
The SQT Program as a Criterion-Referenced Approach to Proficiency Testing.....272

Hosmer, Clark L., Ph.D.
Estimation of the Impartiality of Military Justice.....135

Hoyt, W. G., Butler, A. K., Germas, J. E., Larson, J. T. and Leung, P. W.
Development of CAI Performance Measures: TACFIRE Tactical Data System.....847

Johnson, Cecil D.
The Impact Number of Women in TOandE Units has on Mission Performance.....574

Kantor, Jeffrey E.
AFHRL Research on Utilization of Women in the Air Force.....521

Knerr, Bruce W.
Relationships Between Attitudes Toward Computer-Assisted Instruction and Training Performance.....797

Knerr, Claramae S. and Epstein, Kenneth I.
Sequential Analysis for Individual Proficiency Decisions.....249

Korotkin, Arthur L., Ph.D.
Job Taxonomy/Duty Module Research for Officer Assignment Definition.....111

Lewiski, Raymond L.
The Search for Fairness in Testing.....590

McAreavy, John F., Ph.D.
Achievement Testing for Various Types of Concentrated Management Courses.....1101

McBride, James R.
Computerized Adaptive Testing and its Potential for Military Testing Applications.....461

McFarland, Barry P. and Carpenter, James B.
Characteristics of Self Selected Job Content.....890

McIntosh, Vergil M.
A Computer Assisted Test Construction System.....379

McMullen, Robert L.
Attaining Equitable Personnel Allocations with Supply and Demand Quotas.....126

Masciangelo, Major W. R.
Staff Noncommissioned Officer Task Analysis Study....52

Meiselman, Herbert, Cobean, Nancy, and Waterman, Day
Evaluation of Five Food-Rating-Feedback Cards....1015

Meredith, John B. Jr., and Robinson, Myron A., Ph.D.
Determination of Minimal Acceptable Performance Levels for Criterion-Referenced, Multiple-Choice Tests....611

Mitchell, Major J. L. and DeNisi, Angelo
Another Look at Peer Ratings....150

Moonan, William J., Ph.D.
DYNA: Dynamic Programming Analysis and the Distribution of Effort....389

Osborn, William C. and Campbell, Roy C.
Developing Skill Qualification Tests....276

Pine, Steven M.
Applications of Item Response Theory to the Problem of Test Bias....511

Pulliam, Robert and Mathers, Boyd L.
Simulation of Combat as a Measure of Individual Proficiency....80

Rafacz, Bernard A.
The Selection of an Occupational Counseling Computer Network System for the Navy Recruiting Environment....765

Ree, Malcolm James
The Development of Statistically Parallel Tests by Analysis of Unique Variance....584

Sands, William A.
Prediction of Enlisted Tenure-(Two Years) The POET-2 Model....955

Scanland, Worth, Ph.D., Korb, Lawrence J., Sands, Franklin F., Ph.D.
Ward, Joseph, Rayner, Gail T., Morgan, R. M., and Unna Huh Jeon
The Institutionalization of Change in Military Training....306

Sellman, Major Wayne S. and Gambrell, Captain Melvin B.
Testing of Southeast Asia Refugees to Determine Military Enlistment Eligibility.....865

Seuberlich, Colonel H. E.
Problems of Billets Classification as seen by the Federal Armed Forces Association (DBwV).....632

Shipley, Brian D. Jr., Gerlach, Vernon S., and Hagin, William V.
Measurement of a Pilot's Control Skills in an Instrument Flight Maneuver.....202

Shriver, Dr. Edgar L.
Ratio of Forces Techniques for Scoring Units in Combat Simulations.....1282

Smith, M. J., Ph.D.
Naval Enlisted Advancement System.....7

Snyder, Chaplain (LTC) John S., Jernigan, Chaplain (LTC) D. R. and Lewis, Chaplain (MAJ) Cecil D.
The Chaplain Description Scale: A Peer Evaluation Without Tears.....166

Stein, Earl S.
Job Knowledge Testing as a Measure of Training Effectiveness.....717

Steinheiser, Frederick Jr., and Snyder, C. Wesley Jr.
Score Quality Issues Related to Individual and Weapon Crew Criterion-Referenced Performance Tests.....1166

Symington, Lawrence E.
Worker Job Satisfaction and Opinion in Military Food Service System Design.....1045

Sympson, James B.
Estimation of Latent Trait Status Using Adaptive Testing Procedures.....466

Swanson, Major Ronald G.
Multiple Choice Tests: How Many Alternatives?.....104

Tartell, Joseph S., Captain, Ruck, Hendrick W. and Driskill, W. E., Ph.D.
Job Knowledge Analysis: A New Approach.....706

Vale, C. David
Adaptive Testing and the Problem of Classification.....488

Valentine, Lonnie D. Jr.
Prediction of Technical Training Success from ASVAB and Educational Background.....923

Van Cleve, Roy R.
From Task Inventories to Performance Appraisal Forms.....1265

Vaughn, Captain David S.
Prediction of Test Outline Weights from Occupational Survey Data.....435

Waldkoetter, Raymond O., Ed.D., Gilbert, Arthur C. F., Ph.D.
Hawkins, Helmut H., Wethy, Roy B., Raney, James L., Ph.D.
Organizational Effectiveness Related to Non-Mos Task Performance of a Critical Subgroup..... 1060

Establishing Training Priorities in an Army Supply Military Occupational Specialty.....836

Walker, LCDR C. Lee
Content Validity for Training Analysis (A Practical Application).....577

Weissmuller, Johnny J. and Kauffman, Britt
Data Processing Problems in Collecting Job Survey Information.....783

Wiant, Allen A.
Discussion of a Study of Occupationally Transferable Skills.....616

Wilson, VADM J. B.
Future Trends in Military Training - A Naval Perspective.....1

Wiltshire, Captain Robert B.
Media Applications in the SQT Program.....298

Yellen, Ted M. I.
Computer-Assisted Occupational Counseling: A Prototype Model.....743